2013

Finding differential splice junctions in RNA-Seq data as transcriptional biomarkers for prostate cancer

Ahmad Tavakoli
Universty of Windsor

FINDING DIFFERENTIAL SPLICE JUNCTIONS IN RNA-SEQ DATA AS TRANSCRIPTIONAL BIOMARKERS FOR PROSTATE CANCER

by
Ahmad Tavakoli

A Thesis
Submitted to the Faculty of Graduate Studies
through Computer Science
in Partial Fulfillment of the Requirements for
the Degree of Master of Science the
University of Windsor

Windsor, Ontario, Canada
2013

FINDING DIFFERENTIAL SPLICE JUNCTIONS IN RNA-SEQ DATA AS TRANSCRIPTIONAL BIOMARKERS FOR PROSTATE CANCER

by

Ahmad Tavakoli

APPROVED BY:

L. Porter
Department of Biological Sciences

A. Ngom
School of Computer Science

L. Rueda, Advisor
School of Computer Science

A. Mukhopadhyay, Chair of Defense
School of Computer Science

September 9, 2013

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Alternative RNA splicing is a naturally occurring phenomenon that has been associated with different types of cancer. Detecting splice junctions in the genome of an organism is the key to the study of alternative splicing. RNA-Seq as a high-throughput sequencing technology has recently opened new horizons on the studying of various fields of transcriptomics, such as gene expression, chimeric events and alternative splicing.

In this research, we study prostate cancer from the viewpoint of splicing events as the second most common cancer in North America. We have proposed a method for differentially detecting splice junctions, and in a broader sense splice variants, from RNA-Seq data. We have designed a 2-D peak finding algorithm to combine and remove the dubious junctions across different samples of our population. A scoring mechanism is used to select junctions as features for prediction of cancer RNA-Seq data belonging to patients diagnosed with prostate cancer against benign samples. These junctions could be proposed as potential biomarkers for prostate cancer. We have employed support vector machines which proved to be highly successful in prediction of prostate cancer.

Dedication

To my beloved parents, for all the years of support and encouragement.

Acknowledgements

This research project would have not been possible without the support of many people. The author wishes to express his gratefulness to his supervisor, Dr. Luis Rueda who was abundantly helpful and supportive and offered priceless assistance and guidance. Deepest appreciation are also due to the members of the supervisory committee, Dr. Alioune Ngom and Dr. Lisa Porter for all their support and guidance, and likewise Dr. Dora Cavallo-Medved for her support and guidance. Special thanks also to all his graduate friends, especially group member and a dear friend, Iman Rezaeian for sharing the invaluable assistance.

The author wishes to express his love and gratitude to his family; his loving parents, grandparents, and brothers, for their understanding and endless love through the duration of his studies. The author would also like to convey thanks to all his friends, specially Navid Shakibapour for his support and encouragement.

Contents

List of Figures

List of Tables

Part I

Background

Chapter 1

Introduction

The central dogma of molecular biology is the cornerstone of modern genetics. It consists of two main transformations, DNA becoming RNA, and RNA becoming protein. The first transformation is called transcription, in which from a helix-shaped double-stranded DNA sequence, a single stranded complementary RNA sequence is produced. The resulting RNA can be coding RNA or non-coding RNA such as rRNA, tRNA, snRNA, or lincRNA. The second transformation is called translation, where coding RNA sequences, transcripts, make proteins that are the main operators in a cell [34]. Between these two steps, there is also a process that is called *RNA splicing*, or simply *splicing*. During RNA splicing, parts of the coding gene are removed, which are called *introns*, and other parts are preserved, which are called *exons*. Figure 1.1 which depicts RNA splicing, shows the introns with a dark color and exons using a white color. Introns are removed and exons are retained. The points where introns and exons are separated from each other during splicing, are called *splice junctions*. The result of this process is called messenger RNA (mRNA). Messenger RNA consists of ribonucleotides, which, in turn, code for amino acids and amino acids are the building blocks of proteins.

Splicing does not happen in the same way, all the time for the same gene. Splicing might be altered in different ways, which would lead to the same mRNA with a slight difference such as an included intron or a deleted exon. This process is called *alternative splicing*, which has recently been shown to be more prevalent and influential in gene functions than what was believed before. Although alternative splicing happens as a normal process in eukaryotes, recent studies have revealed that variations in splicing patterns are associated with some diseases such Alzheimer's disease, and alternative splicing also regulates genes that are associated with cancer [5; 59].

It is estimated that 95% of the multiexonic genes in humans are alternatively spliced [43]. As such, alternative splicing explains in part the complexity of mammalians given their small number of genes compared to other organisms [5]. It also makes the synthesis of several different proteins from the same gene possible for higher eukaryotes [1]. Alternative splicing has been observed using different methods, such as exon skipping which is the most common way.

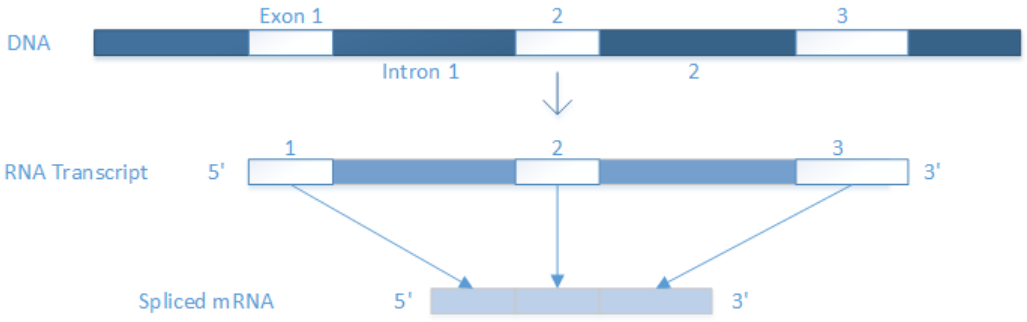

Figure 1.1: Schematic view of RNA splicing.

1.1 Sequencing

In spite of the fact that modern DNA sequencing started in 1977 with the first complete DNA sequence, the first sequencing of DNA happened in 1968, merely fifteen years after the discovery of the double helix [24]. Completion of the Human Genome Project in 2003 was the pinnacle in sequencing, enabling new ways to answer questions in evolution, biology and the environment.

In genetics, the act of determining nucleic acid bases of a DNA or RNA molecule at single base pair resolution in the correct order, is called *sequencing*. Many applied fields, such as medical diagnosis, biotechnology, forensic biology and biological systems have been revolutionized by the knowledge of DNA sequences. Obtaining complete genomes and transcriptomes of numerous types and species of life, including the Human Genome Project, has become possible only by modern sequencing technologies.

High-throughput sequencing, or *next generation sequencing*, has brought down the cost and time of sequencing significantly during recent years. This has happened by parallelizing the sequencing process, leading to the production of millions of sequences concurrently [22]. Sequencing has moved from small research labs that could take months and years for cloning and sequencing of a target gene, to the industrialized large-scale instruments, and with the advent of next generation sequencing to the bench-top instruments in the labs.

RNA-Seq

RNA-Seq is a high-throughput sequencing technology to sequence a cDNA molecule to retrieve genetic information regarding the sample's RNA or transcriptome. RNA-Seq provides single-base resolution and deep coverage and can be used to measure and quantify gene expression levels, study differentially spliced transcripts, non-coding RNAs, small

RNAs, transcriptional structure of genes in terms of their start sites ($3'$ and $5'$ ends), chimeric events, gene fusion, and post-transcriptional modifications.

RNA-Seq aims to provide us with the content of mRNA. However the mRNA molecule, in contrast to DNA, is single stranded. This property makes it unstable and hard to sequence, and as a result it is transformed into cDNA which is double-stranded and stable to be sequenced.

1.2 Biomarkers in Diseases

1.2.1 Biomarkers

A biological marker, or biomarker, has been defined by the National Institutes of Health Biomarkers Definitions Working Group as "a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention". Biomarkers are now widely considered as endpoints in basic and clinical research [56]. A wide variety of events have been considered as biomarkers previously including chimeric events and splice variants, which is the focus of this research [5; 27].

A chimeric event happens when parts of a gene, merges with other coding sequences of another gene that results in formation of a new gene [27]. Although this chromosomal rearrangement process is mostly related to the development of cancer, recent studies suggest that trans-splicing also might occur frequently in healthy cells with regulation [19]. Trans-splicing is a chimeric event in which RNAs that have been formed separately, splice together and form a new RNA.

Based on the observation of $3'$ and $5'$ splice sites at the junction of almost all of the

chimeric RNAs in their research, Kannan et al. [27] suggested that the formation of these chimeras has happened as a result of splicing. This finding implies the close relationship between these two events. From a bioinformatics point of view, splicing and trans-splicing are quite similar. The main difference in the process of their detection is that when looking for possible splicing events, the segments of reads can not be mapped to different genes, while the same is possible for trans-splicing.

1.2.2 Prostate Cancer

Prostate cancer is the most common cancer in North America after skin cancer, which affects 1 out of each 6 men [55]. It is estimated that 238,000 men will be diagnosed in 2013 with prostate cancer, and 30,000 men will die because of it [55]. Adenocarcinoma is the most common type of prostate cancer which occurs when cells lose their natural control over growth, maturation and death.

There have been many studies regarding prostate cancer using RNA-Seq data, which covers a wide range of applications including genome wide association and variation studies, somatic mutations, non-coding RNAs, chimeric RNA and gene fusion [15]. Feng et al. [15] conducted an extensive survey on the most recent alternative splicing studies in cancer using RNA-Seq data. They included a review of recently developed RNA-Seq analysis tools and also a set of publicly available RNA-Seq datasets. Kannan et al. [27] found chimeric RNAs using RNA-Seq data in prostate cancer [27]. They detected 27 previously-unknown highly-recurrent chimeric RNAs.

Pflueger et al. [44] concentrated on the discovery of new gene fusions in human prostate cancer with RNA-Seq data, and detected seven new gene fusions related to prostate cancer including the renowned TMPRSS2-ERG. Xu et al. [68] has studied RNA-Seq data from five

prostate cancer patient and identified variations of chromosomal rearrangements, insertions and deletions [68]. They discovered 92 human genes that had undergone somatic mutations. Xu et al. [68] has also determined from the data that the gene TNFSF10 is unable to induce apoptosis, and as a result it further boosts progression of abnormal tumors. Sahu et al. [51] observed an interesting relationship between RNA splicing and prostate cancer rates among different ethnic groups in America. In this recent study on RNA-seq enrichment of long non-coding RNAs and alternative splicing, prostate cancer has been showed to have a significantly lower incidence rate among Chinese population, and a significantly higher rate among African Americans comparing to Caucasian men [51]. They also noted some previously unknown gene fusions, among them the fusion between the genes TMPRSS2 and ERG.

Wang et al. [64] tackled the problem of discovering differentially spliced genes from two separate RNA-Seq experiments. They designed their solution based on a negative binomial (NB) distribution model for detection of splice junctions. They obtained their data for the RNA-Seq library from human kidney and liver samples, but the method can be applied to other sources, such as prostate cancer samples [64]. Prensner et al. [46] put their focus on non-coding RNAs (ncRNAs) as emerging key molecules in human cancer. They discovered a previously unannotated long intervening non-coding RNA (lincRNA), PCAT-1, that is related to the progression of prostate cancer. In summary, there have been various studies on prostate cancer recently based on RNA-Seq with the focus on chimeric RNAs or gene fusions. These studies mostly used traditional statistical tests and focus on biomarkers at the individual level and/or group of patients.

1.3 Machine Learning

RNA-Seq experiments produce a vast amount of data, which requires significant computational resources in terms of time and space. Machine learning methods have proven to be crucial for data analysis on this scale [13; 50; 69; 8]. These methods provide obvious advantages in terms of accuracy and adaptability and have been extensively used in transcriptomics previously to study cancer [65; 48]. Feature selection, classification, and clustering are among the significant applications of machine learning in bioinformatics.

Support vector machines (SVM) are machine learning methods which have been proposed by Cortes and Vapnik [12] [62] in 1995 based on statistical learning theory, and have since been used extensively on a wide range of applications including bioinformatics. SVM follows a data-driven approach towards solving classification problems. High accuracy and capacity to handle high-dimensional data such as gene expression are among the advantages of using SVM for transcriptome analysis [53].

1.4 Motivation

The detection of biomarkers would have a meaningful impact on diagnosis and treatment of cancer. The validity of using alternative splicing, and splice variants as a biomarker for cancer has been widely studied [5; 54; 42; 17]. Reliable detection of splice junctions is the most important step towards discovery of alternative splicing.

Kannan et al. [27] studied chimeric events on prostate cancer and discovered chimeras that were only present in prostate cancer data. They also concluded that formation of these chimeras are mediated by splicing which led us to the idea of studying differential splice junction detection on prostate cancer. We applied our model on the same RNA-Seq dataset

as Kannan et al. [27]. The innovative aspect of this work is based on using machine learning and pattern recognition techniques for the purpose of classification and feature selection, as well as data integration and processing. Our study also focuses on the discovery of splice junctions as a feature for classification of cancer.

1.5 Problem

The problem that we have addressed in this work is finding splice junctions from RNA-Seq data that could be proposed as biomarkers for prostate cancer. Finding a reliable and accurate way of detecting splice junctions on an RNA-Seq dataset is the first part, and extracting meaning from junctions belonging to 30 different samples of two different classes is the second part of the problem that we tackle in this study.

1.6 Contributions

The main contributions of this thesis are:

- Developing a model for combining and filtering out splice junctions on large scale data using peak-finding in 2-D histograms.
- Designing a method to propose splice junctions as biomarkers based on differential results among cancer and benign samples.
- Development of a system for prostate cancer prediction using the biomarkers and support vector machines.

1.7 Thesis Organization

This thesis comprises six chapters. Chapter II discusses RNA-Seq as an emerging technology in sequencing, its advantages over previous methods and the challenges that researchers face using it. In Chapter III, we provide an introduction to splice junctions which will be followed up by a survey on splice junction discovery methods. Next chapter includes the methods and materials that we have used to address our problem at hand. Chapter V includes the results of the experiments that we conducted, as well as comparisons and discussions regarding them. Finally, in Chapter VI, a conclusion on this topic is made and future works are discussed.

Chapter 2

RNA-Seq

2.1 RNA-Seq Technology

RNA-Seq has been dubbed revolutionary by the scientific community because of its ability to transform our knowledge about eukaryotic transcriptomics at a level of detail and precision that has never been studied before [22]. Next-generation or high-throughput sequencing provides a way to sequence cDNA to study a sample's mature RNA sequence, which is called RNA sequencing or RNA-Seq. Next-generation sequencing supplies a ground for massive transcript expression analysis, and has become the prominent approach to study transcriptomics since 2008. Short reads acquired from high-throughput sequencing technologies can be used for studying transcriptome and gene structure identification. RNA-Seq can be used for cellular phenotyping and help establishing the etiology of diseases characterized by abnormal splicing patterns.

Before the invention of RNA-Seq, microarrays were the way to study the transcriptome. The main methods for this purpose are hybridization-or sequence-based approaches [66]. Limited dynamic range resulting from high level of background, saturation signals

and being dependent on existing genome annotations are amongst the main restraints of microarrays. The ability to study and measure the transcriptome without prior knowledge of the reference genome, is one of the advantages of RNA-Seq over microarrays. This enables researchers to detect previously unknown transcripts. RNA-Seq is also more sensitive in detecting changes in low expressed transcripts [70].

Detection of chimeric transcripts and gene fusion is among many RNA sequencing applications that are being studied extensively [27; 33; 38; 37; 45]. Chimeric RNAs have been suggested to be a possible biomarker in at least two recent studies [49; 36]. Kannan et al. [27] used paired read information to search for chimeric events across genome. They looked for paired reads that could be mapped to a different gene either in the genome or transcriptome. They used different filtering strategies to reduce the number of false positives. The number of mismatches in the initial mapping of the reads is a criteria for this purpose, which has been set to a tolerance rate of two mismatches.

RNA-Seq is the favorite approach to study gene expression at a base-level with high coverage, it supplies enough reads for the purpose of detecting alternative splicing. One of the keys to profiling this genetic information is the identification of intron-exon boundaries or splice junctions. In RNA-Seq, the exact nature of splicing events is buried in the reads that span intron-exon boundaries. The accurate and efficient mapping of these reads to the reference genome over these boundaries is a major challenge, which is a requirement for studying RNA splicing.

2.2 RNA-Seq Preprocessing

2.2.1 Preparation

There are various technical approaches for preparation of an RNA-Seq experiment. The first step that is common among all technology platforms is determining the amount of RNA that is required. This amount could be different based on the sequencing platform and priming method [67]. The majority of RNAs ($>90\%$) existing in cells are ribosomal RNA (rRNA). The remaining RNAs are composed of mRNA and other types of RNAs. As a result, they do not provide useful information regarding the transcriptome. There are various techniques to concentrate the sequencing on non-ribosomal RNAs. Selective enrichment of mRNA is a method that can be carried out by enriching the PolyA tail present in mRNA molecules.

The resulting mRNA from the enrichment process should undergo the priming process in the next step. This could be done using either random primers or oligo-dT primers [15; 67], which is also called mRNA-Seq. Another consideration that should be taken into account, especially for comprehensive RNA-Seq experiments in organisms like human and mouse with complicated genomes, is creating double-stranded cDNA to maintain strand specific information of the RNA [67].

2.2.2 Paired-end Reads

RNA-Seq reads comes in two types, single-end reads and paired-end reads. Paired-end reads which are being used more and more in transcriptomic studies, consist of two fragments obtained from both ends of a DNA fragment [66]. The length of an RNA-Seq read could be up to $500bp$ depending on the sequencing conditions. Because of the limitations of the technology, only sequences from the tails of that read can be obtained. In case of

obtaining both ends of the read, it is paired-end sequencing or it is single-end if we have only one of the reads.

Paired-end sequencing provides us with many advantages in data analysis, also it requires no more DNA as single-end sequencing and hence is more efficient [25]. Using Illumina technology, end users are able to choose their required insert size between forward and reverse strands of DNA. A dataset of 200 million reads of 2 x 75 bp is the typical result for a paired-end sequencing run using Illumina technology. An insert size between 120 bp and 170 bp is generally suitable [67].

Using single-end sequencing, only one strand of the DNA fragment is sequenced and the information from the other strand is lost. Being able to sequence the other strand, gives us the capacity to do the aligning more accurately and reduces the number of errors.

Given the insert size of each read and position where one of the reads map to the reference genome, we know the direction and approximate position of the other read. PASSion [70], one of the methods studied, use this information to optimize its algorithm's performance.

2.3 RNA-Seq Data Analysis

In RNA-Seq, bioinformatics faces similar challenges as other high-throughput sequencing technologies. The main challenges are the development of algorithms and tools for storage, retrieval and processing of large datasets containing the information related to millions of short reads for each RNA-Seq run. The efficiency of these methods becomes more critical in dealing with low abundance transcripts where the error rate becomes higher. However, the same problem existed for previous technologies such as microarrays.

2.3.1 Coverage

Depth of sequencing is another important aspect of any sequencing technology and also RNA-Seq. Sequencing coverage, or the percentage of transcripts being read, is affected by the depth of sequencing. Generally speaking, more sequencing depth will lead to higher coverage. Sequencing depth, in turn is related directly to the cost of the experiment [26; 66]. To sequence simpler transcripts, for which alternative splicing has not been observed in them, lower depths are sufficient. However, higher depth might be needed in specific cases, such as when we are investigating rare events, as in lowly-expressed transcripts. The reason is that in RNA-Seq different transcripts are expressed depending on their gene expression levels. Also higher depth might lead to more statistically significant results.

2.3.2 Mapping

Almost in any RNA-Seq data analysis, mapping the reads is the first step to perform. RNA-Seq is a high-throughput sequencing technology, and as such, it provides us with numerous short reads. Depending on the details of the sequencing platform and mapping technology being used, short reads can be mapped directly to the reference genome or can be assembled to form contigs. These contigs can be used to reveal the transcriptome structure [66].

To maintain a high standard for the reads that are being mapped, and decrease the chance of errors and dubious results, most of the mapping algorithms implement one or several ways to filter out reads that have a given number of base pairs with quality scores lower than a particular threshold. This generally also increases the speed of subsequent processing.

For large transcriptomes, as we have both a high number of reads and lengths of reads are short, some reads may align to multiple places across the reference genome. This alignment could have happened with different alignment scores computed by mismatches,

deletions, and insertions. Based on the application, this problem can be addressed by discarding reads that map to more than a defined number of places in the genome, or filtering out the reads based on their corresponding alignment scores. The next step in most RNA-Seq experiment pipelines is on selection of unique reads and removal of repeated reads.

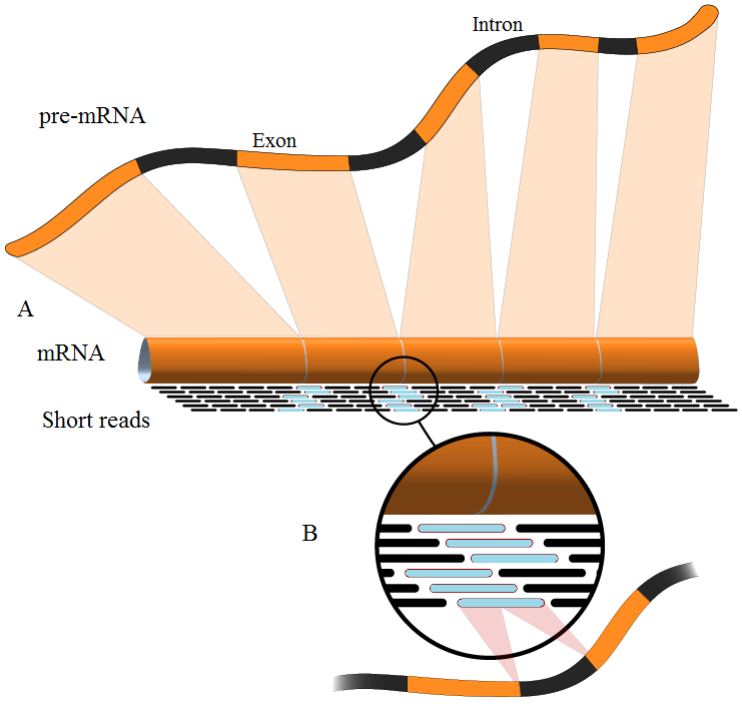

Figure 2.1: Alignment of RNA-Seq reads across splice junctions [61]. Courtesy of User:rcogs, Wikimedia Foundation.

Part A of Figure 2.1 illustrates the process of splicing. In Part B, the process of mapping back RNA-Seq reads to the reference genome is shown, which leads to the detection of splice junctions. The reads that span a splice junction are shown in a light color, and the reads that map to a single exon are shown in black. This figure demonstrates that light-colored reads are split when aligning back to reference genome.

Before introducing alignment solutions that are designed and developed specially to

map short reads in higher volumes, it was only possible to map 10 million short reads using BLAST or BLAT, which needed 6 and 43 hours to complete respectively (see Li et al. [31]). Considering the fact that some sequencing platforms are capable of generating 200 million reads in a single run, it could be concluded that the application of common read aligners is not feasible with next-generation sequencing technologies. To address this issue, many short read aligner programs such as ELAND by Mortazavi et al. [39], SOAP and SOAP2 by Li et al. [31] [32], BOWTIE by Langmead et al. [29], SMALT [70], and BWA by Li and Durbin [30] were developed to facilitate RNA-Seq analyses.

2.4 RNA-Seq Datasets

Studies on RNA-Seq could be conducted on either real or simulated datasets. Simulated datasets give us the ability to form reads *in silico*, and hence having the option to try our methods for different environment-dependent variables such as read length and quality, insert size, expression level, etc. Also, it provides us a reliable means to compare different methods against each other.

Most of RNA-Seq datasets are made publicly available for further studies. Various file formats have been proposed for storing next-generation sequencing files. Sequence Read Archives (SRA) format is designed for storage of large amounts of sequence data, for this reason the data is compressed and not easily read. SRA format is one of the standard formats used by major genomic databases around the world including NCBI, EBI, and DDBJ to store sequence data [41]. Figure 2.2 shows a paired-end RNA-Seq read in FASTQ format from the dataset that we use in this study. Each read is represented in 4 lines in this figure. First and third lines are read IDs, second line includes the read sequence and the fourth line is the quality score. As can be seen, the read ID and the read length for both

strands is the same. However, read sequence and quality scores are totally different.

The dataset that we use in our study contains more than half a million paired-end RNA-Seq reads that have been acquired using Illumina Genome Analyzer II platform [27]. While this dataset occupies more than 16*GB* when stored in SRA format, storing it in FASTQ format takes 5 times the amount of physical memory.

```
@SRR057653.1 HWUSI-EAS230-R:2:1:27:501 length=36
AAAAAATATGGTTAAAAACTGTATANANNANNNNNT
+SRR057653.1 HWUSI-EAS230-R:2:1:27:501 length=36
=6=@8:><)8=+-B>=:6?######!#!!#!!!!!#


@SRR057653.1 HWUSI-EAS230-R:2:1:27:501 length=36
CTTTAATACACATTAAGTCATTTAATTCTCACCTAG
+SRR057653.1 HWUSI-EAS230-R:2:1:27:501 length=36
@0=+@@@B@?>@:/@B='><%'7>A908@<B@B==5
```

Figure 2.2: Sample RNA-Seq read in the FASTQ format.

Following its introduction to the scientific community, RNA-Seq has soon found its place in research in various fields such as gene expression profiling and RNA splicing events, and started a new era in transcriptomics. The data acquired from RNA-Seq is comprehensive in nature and has shaken the field of transcript identification, and has contributed significantly to the process of transcriptome assembly. However, the challenge remains in the bioinformatics field to develop algorithms for analyzing these data and extracting biological meaning from them.

Chapter 3

Splice Junction Detection

Detecting splice junctions has always been one of the interesting fields of studying transcriptomics, and microarrays have been used extensively for this purpose in the past. Since 2008, next-generation sequencing has become the prominent method to study transcriptomics. In this chapter, we present the major works dealing with the detection of splice junctions using RNA-Seq data.

All the reviewed works have been categorized according to the method that the authors have developed to detect splice sites. In the first category, we present the methods developed by De Bona et al. [13], Dimon et al. [14], and Lou et al. [35]. The authors in this section have used some sort of machine learning algorithms for detection of splice junctions in their approach. In the second section of this literature review, we would study the methods which try to assess the reliability of a possible splice junction by a read-counting method. This section consists of works by Trapnell et al. [58], Wang et al. [63], Huang et al. [23]. PASSion designed by Zhang et al. [70] also belongs to this group. The third section focuses on works that use other methods as their main way of removing false positives from their results. These papers include those of Au et al. [3], Ameur et al. [2], and Bryant et al. [7].

3.1 Alternative Splicing

Understanding the nature of alternative splicing is very important in the detection of splice junctions. Figure 3.1 illustrates examples of alternative splicing. In this figure, exons and introns are represented by rectangles and lines respectively. The most prevalent way that alternative splicing happens is exon skipping which is shown as number 4 in the figure. Exon skipping happens when an exon that is supposed to be part of the mRNA is removed. Intron retention (Number 5) is the opposite, when an intron is kept in the mRNA, as it was supposed to be removed. Alternative $3'$ and $5'$ splice site usage changes could also happen, which will lead to different splicing patterns (Numbers 1 and 2). Splice sites are shown with dark boxes in the figure. Alternative promoters and alternative poly(A) sites are also other forms of alternative splicing [43; 5] (Numbers 6 and 7).

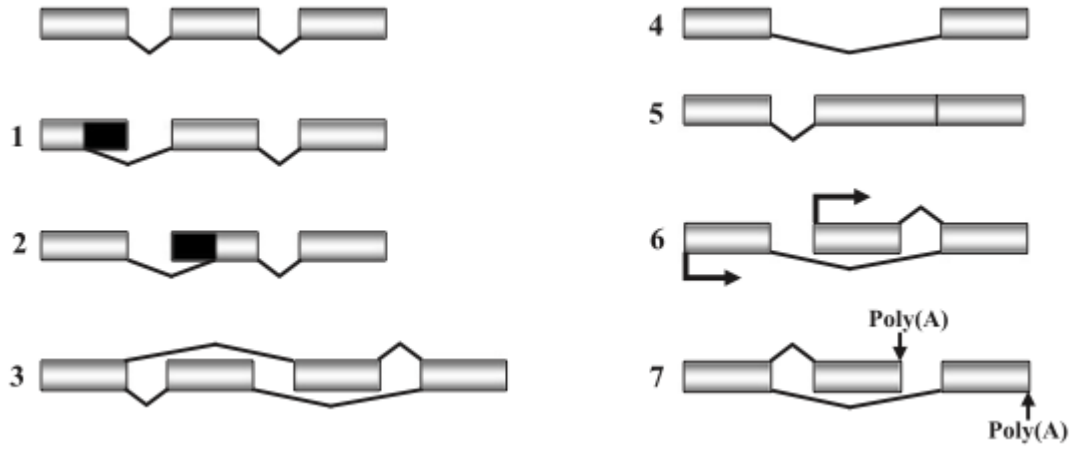

Figure 3.1: Different types of alternative splicing [5]. The top leftmost image shows the normal way of splicing where introns are removed and exons are retained.Courtesy of Elsevier.

3.2 Methods for Splice Junction Detection

RNA-Seq rose to prominence in research after 2008, and as a result, all methods of this review have been published over a time spanning less than five years. Researchers may not have enough time to study the works of their peers and make comparisons and experiments on already developed methods. The works by Ameur et al. [2], Au et al. [3], and Bryant et al. [7] which are studied in this review, have been published within a month. Obviously, none of them had the chance to refer to each other. The interesting point is that these papers fell into the same section in this review, which implies the similarity of their work.

3.2.1 Methods based on Machine Learning

In this section, we review the methods developed by De Bona et al. [13], Dimon et al. [14], and Lou et al. [35]. All these methods apply some sort of machine learning technique in their pipeline. QPALma developed by De Bona et al. [13], uses SVM and HMMSplice by Dimon et al. [14] utilizes hidden Markov models. Lou et al. [35] use maximum likelihood estimation in their approach to splice junction detection. Table 3.1 reviews the availability and update frequency of these methods as software packages.

Methods based on Support Vector Machines

Short reads acquired from high-throughput sequencing technologies can be used for studying the transcriptome and gene structure identification. Aligning these reads over intron/exon boundaries is a requirement for this purpose. De Bona et al. [13] do not refer to any previous work which relates to the subject of this review.

De Bona et al. [13] developed their approach, called QPALMA, in three independent parts, splice site prediction model, a dynamic programming algorithm and a scoring func-

tion. QPALMA aims to align short reads to the reference genome; splice site prediction helps this approach to achieve better results. This part is based on a machine learning approach which uses a set of donor and acceptor sites to train a SVM predictor. The authors propose three different extensions for the Smith-Waterman algorithm for aligning the reads to the reference genome.

The authors state that they trained their algorithm using a simulated dataset of 10,000 previously aligned sequences. The alignment error rate for these rates has been calculated by incorporating different available pieces of information. The authors also tried their approach on a dataset of spliced and unspliced 2.98 million reads of forward strands of chromosome 1.

De Bona et al. [13] claim that they could align 10,000 *in silico* spliced reads with an error rate of 1.78% incorporating quality information, intron length model and splice site predictions. The authors state that it was the best rate that they could have achieved. The authors also claim that QPALMA aligned spliced and unspliced reads with a 5.2% and 1.2% error rates respectively.

The authors claim that they could successfully exploit all information sources to align short reads over exon boundaries. The authors claim that their approach works reasonably well for all next-generation sequencing platforms, including *Illumina* sequencing, which has been tried in their experiment. The authors state that their method can be extended to exploit homo-polymer errors, which is available for Roche's 454 sequencing platform.

Method based on Hidden Markov Models

During the past decade, there has been a growing appreciation of the importance of alternative splicing as a mechanism for organisms to increase proteomic diversity and regulatory

complexity. According to Dimon et al. [14], the ability to detect alternative splice isoforms with accuracy and sensitivity is the key to comprehensive RNA-Seq analysis. The authors refer to previous work by Mortazavi et al. [39], Trapnell et al. [58], Bryant et al. [7], and Ameur et al. [2].

They note that the method developed by Mortazavi et al. [39] does not address the question of novel junctions and cannot be used for organisms with incomplete or inaccurate genome annotations. They state that the algorithm developed by Trapnell et al. [58] performs best on mammalian transcripts with relatively high abundance, but has defects in more compact genomes and with non-canonical junctions. They note that the method proposed by Ameur et al. [2] has the requirement for at least one read to split evenly across the exon-exon boundary which reduces sensitivity in low coverage datasets and transcripts. Also, they claim that this method supports only ABI SOLiD reads.

Dimon et al. [14] state that SuperSplat, the method developed by Bryant et al. [7], requires both pieces of a read to be exact matches to the reference sequence and conclude that it is not robust against sequencing errors or single-nucleotide polymorphisms. The authors claim that the algorithm designed by Au et al. [3] considers only canonical splice junctions and requires read lengths of $50bp$ or greater.

Dimon et al. [14] claim that they have developed a method to avoid the inherent bias introduced by relying upon previously defined biological information. Their algorithm, called HMMSPlicer, works by dividing each read in two halves and seeding the read-halves against the genome and using a Hidden Markov Model to determine the exon boundary. They claim that both canonical and non-canonical junctions are reported and a score is assigned to each junction, which is dependent on the strength of the alignment and the number and quality of bases supporting the splice junction.

The authors claim that they compared their algorithm with those designed by Trapnell et al. [58] and Au et al. [3]. They state that they analyzed the performance of their method on simulated reads and three publicly available experimental datasets.

Dimon et al. [14]. include the detailed results of their experiments with different algorithm parameters on the examined datasets. The authors state that in comparison with TopHat, HMMSplicer shows its ability to find more junctions with a similar level of specificity in each of the tested datasets. They state that in comparison with SpliceMap by Au et al. [3], their method achieves 7% more matching junctions for human datasets, and it outperforms SpliceMap in the low sequence quality *A. thaliana* dataset.

The authors claim that HMMSplicer combines high sensitivity with a low false positive rate, functions properly on datasets with low quality sequence reads, performs well in datasets with uneven coverage, identifies many junctions in low abundance transcripts and also identifies non-canonical junctions. It also finds true novel junctions in genomes with incomplete annotation. The authors claim that their algorithm is the only software package that provides a score for each junction, reflecting the strength of the junction prediction.

Method based on Maximum Likelihood Estimation

Studying the way that alternative splicing affects a biological system is as important as studying its fundamental regulatory mechanisms, and as RNA-Seq provides the ability to analyze the transcriptome in a base-level resolution and high coverage. Lou et al. [35] refer to the work by Mortazavi et al. [39], Trapnell et al. [58], Ameur et al. [2], Au et al. [3], Bryant et al. [7].

The authors state that the approach presented by Trapnell et al. [58] depends on the canonical splice site motifs. They also state that as the methods developed by Ameur et al.

[2], Au et al. [3] and Bryant et al. [7] designed based on the idea of read counting, sequencing depth can significantly affect their performance.

Lou et al. [35] proposed an approach based on maximum likelihood estimation, which relies on geometric-tail distribution of intron lengths for aligning of paired-end RNA-Seq reads. The authors used a package named ABMapper, which was particularly developed for spliced mapping by the same team as the authors and is explained in Lou et al. [35]. They state that their approach is an empirical probabilistic model which adopted a two-part distribution, an arbitrary length distribution and a geometric distribution. This method uses maximum likelihood to estimate the most probable location for a paired-end read based on this two-part distribution. The authors stated that their approach works in three models, one without any *a priori* knowledge, and two with expression level and junction-site frequency as *a priori* knowledge.

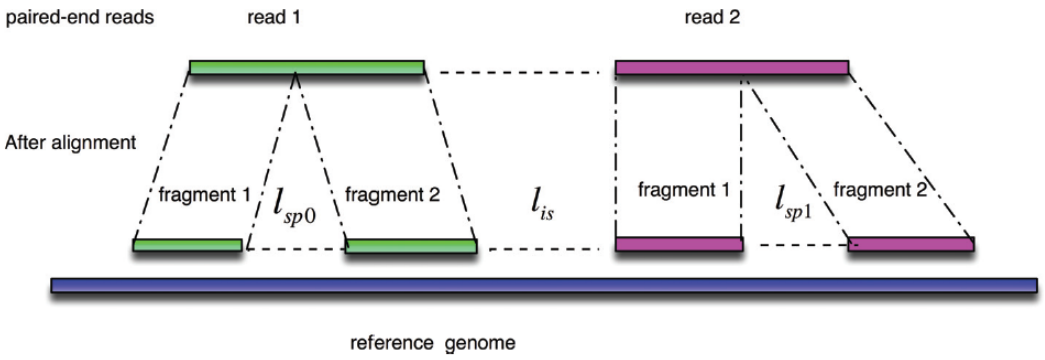

Figure 3.2: Alignment of a paired-end read by the work of Lou et al. [35]. This methodology supports mapping of both reads across splice sites. Courtesy of Biomed Central.

The authors state that they compared their model with methods developed by Trapnell et al. [58] and Au et al. [3]. They used two human lymphoblastoid cell-line datasets for testing purposes. The dataset consist of 8.4 million 75 bp paired-end reads with an approx-

imately 250 bp insert size. The results were validated with the Alternative Splicing and Transcript Diversity (ASTD) database and the Human EST database.

Lou et al. [35] claim that their method could report 53% and 49% more splice junctions compared to the methods by Au et al. [3] and Trapnell et al. [58]. The authors also claim that 60% of the junctions which were predicted only by their method could be validated by the ASTD, which comprised 22% of the total reported splice junctions. According to the authors, this implies that the methods proposed by Au et al. [3] and Trapnell et al. [58] missed at least one-fifth of the true splice junctions. The authors claim that by performing an exhaustive search for junctions in the Human EST database, they found that their method predicted splice junctions with an accuracy of 96%.

The authors claim that their proposed approach can detect 50% more splice junctions than other existing tools. Lou et al. [35] claimed that the reason for superiority of their approach is in using first, ABMapper, which has a much higher sensitivity in spliced-mapping than other approaches, and the second is the geometric-tail based model.

3.2.2 Counting-Based Filtering Methods

The papers reviewed in this section, use filtering methods based on counting the number of reads covering the reference genome. The papers presented by Au et al. [3] and Ameur et al. [2] map the reads by splitting them. The method developed by Bryant et al. [7] use also empirical data for supporting possible junctions. The availability and update frequency of the packages reviewed in this section are shown in Table 3.2.

Year	Author	Package Name	Notes
2008	De Bona et al.	QPALMA	Cited by Trapnell et al. [58], Bryant et al. [7],Wang et al. [63], and Huang et al. [23] Last updated in December 2010. The package is freely available.
2010	Dimon et al. [14]	HMMSplicer	Cited by Huang et al. [23], and Zhang et al. [70] Last updated in November 2010. The package is freely available.
2011	Lou et al.	N/A	Not Cited. The package is not available.

Table 3.1: Availability and update frequency of the software packages mentioned in this section.

Methods based on Mapping reads by splitting

High-throughput sequencing of mRNA opens extraordinary opportunities to identify the spectrum of splice events in a sample on a global scale. The works presented in this section describe the methods developed by Au et al. [3] and Ameur et al. [2] to address this problem. Both of these methods, split the reads initially, and then try to map the fragments onto the reference genome.

Definite fusion transcripts are commonly produced by cancer cells, and detection of fusion transcripts is part of routine diagnostics of certain cancer types. Abnormal RNA splicing is associated with many human diseases. For this reason, methods to identify and quantify splicing events are important in biology and medicine. Ameur et al. [2] refer to the work by Trapnell et al. [58].

The authors state that in the method designed by Trapnell et al. [58], a substantial number of true splice junctions, including junctions with long introns or non-canonical splice sites are outside of the detection range, and also this method is computationally challenging

for transcripts expressed at lower levels.

The authors state that their method consists of a combination of a split-read alignment and the novel SplitSeek program. The alignment is performed using the AB/SOLiD whole-transcriptome-alignment software. The method proposed by Ameur et al. [2], SplitSeek, was developed in a way to find junction reads in which as few as five bases overlap with the other exon. It finds exon-exon boundaries that are supported by several split reads. It is required that each junction be covered by at least two reads with unique starting points.

The authors state that they evaluated their method using public RNA-seq data from single mouse oocytes, which was performed on two independent samples, and consist of 50-bp reads. They also state that they selected 22 base pairs for the anchor length according to the highest number of uniquely mapped split reads that was obtained for this length. Ameur et al. [2] present the results of their experiments in terms of the number of splice junctions and insertions, number of predicted small insertions and deletions within RefSeq exons, and number of predicted splice junctions as a function of the total number of processed reads.

The authors claim that the exon-exon boundaries are identified almost at nucleotide resolution and with a low false-positive rate, less than one in 10,000, for junctions within 100 kb. Ameur et al. [2] state that their method makes it possible to study splice junctions and fusion genes while also measuring the gene expression using RNA-Seq data. They claim, according to their results, that their proposed algorithm has a very low false-positive rate, and they state that acquired false discovery rate of less than one for $1,000$ junctions within $1Mb$ and less than $1/10,000$ for those within $100kb$, is supporting their claim.

Au et al. [3], in their research paper, state that the method developed by Mortazavi et al. [39] is dependent on an annotated exon library and since the exon library is incomplete,

this method cannot find junctions that involve novel splicing events. The authors do not mention any shortcomings of the method presented by Trapnell et al. [58].

Au et al. [3] present their method, SpliceMap, based on the idea of the mapping of half-reads as a way to identify the approximate location of a junction. SpliceMap works in four steps, half-read mapping, seeding selection, junction search and paired-end filtering. It maps both halves of the read to the reference genome by a short read mapping tool.

The authors state that they compared their method with the method described by Trapnell et al. [58] on an RNA-Seq dataset of $23,412,226$ reads. They claimed that they assessed their method's specificity by aligning detected junctions to human ESTs in GenBank. They also stated that they investigated the novelty of discovered junctions by PCR experiments. Au et al. [3] describe a comparison with ERANGE proposed by Mortazavi et al. [39]. They also stated that they compared their method with BLAT, which is a common tool for EST sequences alignment. The authors claimed that they calculated the performance of their method in a specific CPU running time and compared it with TopHat by Trapnell et al. [58].

Au et al. [3] claim that 87.9% of junctions found by their method were supported by EST evidence. They state that SpliceMap achieves more than 95% sensitivity for highly expressed genes, more than 90% for genes with medium expression and $(40-67\%)$ for genes with low expression. They state that more genes detected by SpliceMap are of higher degree $(80-100\%)$ of completeness in junction discovery. Au et al. [3] claim that in a random sample experiment, 85% of novel junctions were validated using PCR experiment.

The authors stated that ERANGE found 160,899 junctions and SpliceMap found 151,317 junctions, among those found by Au et al. [3] method, 23,020 junctions, which were not found by ERANGE, were novel. They also claimed that the BLAT package, achieved a

similar but still slightly lower level of specificity with a much lower sensitivity (70% lower) as compared to SpliceMap. The authors stated that it took 66 CPU hours for SpliceMap and 12 CPU hours for ERANGE to process the data set.

They claim that based on their results, SpliceMap detects more annotated junctions than TopHat, method presented by Trapnell et al. [58]. They claim that $50bp$ reads can support an approach of direct *de novo* detection of splice junctions without the need to first cluster reads to identify accepted exons, and that this approach can achieve significantly higher sensitivity in junction detection than current leading methods of RNA-Seq analysis. They also claim that paired-read information can help reduce false discoveries.

Year	Author	Package Name	Notes
2010	Au et al.	SpliceMap	Cited by Wang et al. [63], Dimon et al. [14],Lou et al. [35], and Huang et al. [23] Last update in October 2010. Source code is freely available.
2010	Ameur et al.	SplitSeek	Cited by Dimon et al. [14], and Lou et al. [35] Not available for download. Not being maintained.
2010	Bryant et al.	Supersplat	Cited by Dimon et al. [14], Lou et al. [35],Huang et al. [23] Not available for download. Not being maintained.

Table 3.2: Availability and update frequency of the software packages mentioned in this section.

Methods based on empirical RNA-Seq data

Next-generation sequencing provides a ground for massive transcript expression analysis. RNA-Seq supplies enough reads for this purpose, and the key to profiling this genetic information is identification of intron/exon boundaries or splice junctions. Bryant et al. [7] refer to the work by De Bona et al. [13], Trapnell et al. [58], and Filichkin et al. [16] in their paper.

Bryant et al. [7] state that the method developed by De Bona et al. [13] relies on the previously known splice sites for training the algorithm which influences the results. Furthermore, they note that QPALMA scores junctions that conform with canonical splicing motifs higher, so it may be inefficient in finding non-canonical splice junctions. The authors state that this problem also applies to TopHat, the method developed by Trapnell et al. [58]. Bryant et al. [7] state that TopHat needs a high number of RNA-Seq reads to build exon islands.

Bryant et al. [7] introduce a new approach, called Supersplat, that uses a hash table as a way to save system memory by storing read sequences as keys and their frequencies as the value. Supersplat uses two parameters to limit the maximum and minimum length of the sequence. Later on, Supersplat builds location indexes based on these parameters. After indexing the reference sequence, Supersplat identifies reads that can be aligned against the reference genome, as possible splice junctions in an iterative process. Potential splice junctions are filtered based on the number of overlapping reads on two intron boundaries.

The authors state that they tested the performance of their method on a set of 3,690,882 *Arabidopsis thaliana* reads. They used the TAIR8 database of annotated junctions to evaluate Supersplat's performance. Bryant et al. [7] also state that they assessed their approach for *de novo* splice junction discovery on a dataset of *Brachypodium distachyon*.

The authors claim that they confirmed 91% of canonical and 86% non-canonical splice junctions using PCR and Sanger sequencing. Bryant et al. [7] claimed that they achieved a predicted positive rate (PPV) of 70% with the minimum read length of 6 and a 90% rate by increasing it. According to Bryant et al. [7], this rate reaches 97% by setting the overlapping number of reads filter to 21.

Bryant et al. [7] claim that their approach is unbiased and exhaustive, but it may generate output files with up to tens of gigabytes in size, and the user should account for determining the befitting criterion to filter out spurious output. The authors claim that the exhaustive approach of their method can discover many previously unknown splice junctions.

3.2.3 Non-counting Filtering Methods

The common element between the papers presented in this section is using various filtering techniques to omit spurious splice junctions. We tried to be as specific as possible in categorizing the papers presented in this section. As all of these methods use simple search methods to find junctions, they need some sort of filtering to detect false positives and gain higher sensitivity. Although, they developed methods that used various strategies for filtering and also were highly similar to each other in nature. The work developed by Trapnell et al. [58] has been categorized as a filtering based on average read depth coverage method, the work of Wang et al. [63] as a filtering on minimum anchor length method, and the works of Huang et al. [23] and Zhang et al. [70] as a filtering using paired-end information and read coverage methods. The availability and update frequency of these methods is shown in Table 3.3.

Methods based on Filtering that use average read depth coverage

Alternative splicing is a significant process in normal cellular functions and also in human diseases. Finding novel splice junctions is an important part of studying alternative splicing. Trapnell et al. [58] refer to previous work by De Bona et al. [13] and Mortazavi et al. [39].

The authors mention two shortcomings of the work of De Bona et al. [13], the first is that their method, QPALMA, depends on a set of known splice junctions from the reference genome and cannot identify novel junctions. They state that the other shortcoming is that De Bona et al. [13] use Vmatch, an alignment program which is not designed to map short reads on machines with small main memories and is considerably slower than other short-read mappers. Trapnell et al. [58] state that ERANGE, the method developed by Mortazavi et. al, depends on available annotation of exon-exon junctions for its main objective, which is gene expression quantification in mammalian RNA-Seq projects.

The authors introduce a new system called TopHat, which works in two phases to find junctions. In the first phase, all reads are mapped to the reference genome using Bowtie, all reads that do not map to the reference genome are set aside as initially unmapped reads. Then, an initial consensus of mapped regions, called exon islands, is computed using the assembly module in a package named Maq. Sequences flanking potential donor/acceptor splice sites within neighboring regions are joined to form prospective splice junctions. For each splice junction, TopHat searches the initially unmapped reads in order to find reads that span junctions using a seed-and-extend strategy. Figure 3.3 shows the pipeline of TopHat, that has influenced splice junction detection methods which were developed after it significantly. This figure shows different stages of splice junction discovery, including mapping against the reference genome, generating exon islands, and mapping initially unmapped reads to the splice sites.

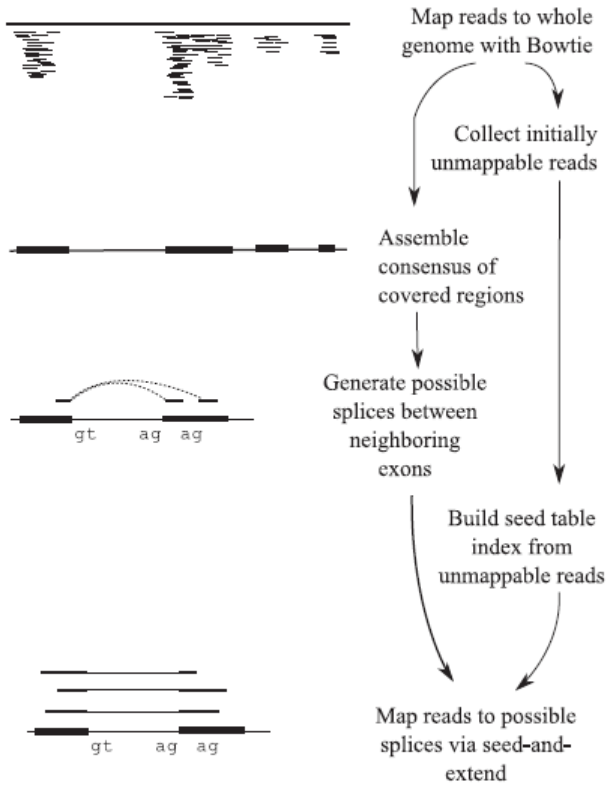

Figure 3.3: The Tophat pipeline which describes necessary steps toward detection of splice junctions [58]. Courtesy of Oxford Journals.

Trapnell et al. [58] state that they conducted an experiment on 47,781,892 short reads using their method, TopHat, and a previously developed method called ERANGE by Mortazavi et al. [39]. The authors claim that their method could discover around 72% of splice junctions compared to annotation-based analysis done by Mortazavi et al. [39] in fewer transcribed regions and 80% of junctions in more actively transcribed regions. They claimed that out of 19,722 newly discovered junctions that they found in their experiment, many of them are true splices, but it is difficult to asses exactly how many of them are genuine.

Trapnell et al. [58] claim that the significance of their work is in its ability to detect novel splice junctions. They also claimed that their tool represents a significant advance

over previous RNA-Seq splice detection methods.

Methods based on Filtering that uses minimum anchor length

Accurate identification and quantification of transcript isoforms is crucial to characterize alternative splicing among different cell types. In addition, sequence variants found within splice sites or splicing enhancer sequences may have functional consequences on alternative splicing. A large proportion of human genetic disorders results from splicing variants.Wang et al. [63] refer to the work by De Bona et al. [13], Trapnell et al. [58] and Au et al. [3]. The authors note that the output generated by the method of Au et al. [3] does not include tag alignments, and hence is incomplete. They do not state any shortcoming regarding the works of others.

According to Wang et al. [63], their method operates in two phases. In the first phase, that is called tag alignment, candidate alignments of the mRNA tags to the reference genome are determined. A set of candidate alignments are computed for each tag as multiple possible alignments may be found for each read, which is shown in details in Figure 3.4. MapSplice uses a double-anchor search method to look for the splice junction. In the second phase which is called splice inference phase, splice junctions that appear in the alignments of one or more tags are analyzed to determine a splice significance score based on the quality and diversity of alignments that include the splice. The most likely alignment for each tag is chosen based on the splice significance score.

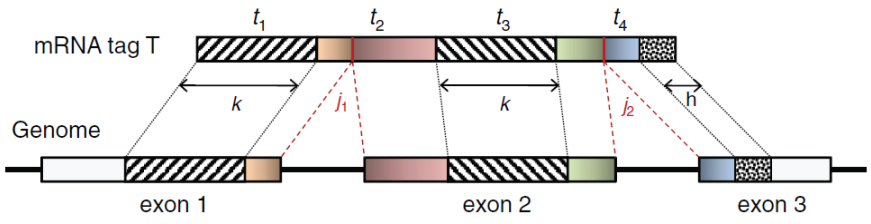

Figure 3.4: Mapping of a read across two splice junctions by MapSplice [63]. Courtesy of Oxford Journals.

The authors state that they evaluated specificity and sensitivity of their method using an experiment on a generated synthetic dataset. They also state that they validated their method using quantitative RT-PCR experiment. Wang et al. [63] state that they achieved a true-positive rate of 96% and false-positive rate of 8% for their method. They stated that over 77% of canonical junctions found by their method were confirmed by known transcripts in GenBank, which was between 6% to 11% higher in comparison by TopHat method by Trapnell et al. [58].

The authors claim that both TopHat by Trapnell et al. [58] and their method were more memory efficient and much faster in experiments than SpliceMap by Au et al. [3]. They also claim that their method performed best by detecting more true-positive junctions and fewer false-positive junctions than the other two methods. They state that longer tags improve both the sensitivity and the specificity of the junction discovery in their method and as well in the method by Trapnell et al. [58]. They claim that in comparison, their method has a higher sensitivity in different tag lengths. They also claim that using read lengths of 75 or $100bp$ yields significantly better sensitivity and specificity for splice detection.

Methods based on Filtering that use paired-end information and read coverage

Splice junction detection is the first step of studying alternative splicing. Alternative splicing is highly effective on diversity of proteins, as it causes different mRNAs to be produced from the same gene. These different mRNAs translate into different protein isoforms. Huang et al. [23] refer to the previous work by Mortazavi et al. [39], De Bona et al. [13], Trapnell et al. [58], Bryant et al. [7], Au et al. [3], Wang et al. [63], and Dimon et al. [14].

The authors state that QPALMA, the method developed by De Bona et al. [13] which uses a machine learning approach, is biased toward splice junctions that are similar to the ones in the training data set. Huang et al. [23] state that low sequencing depth affects the performance of the algorithm developed by Trapnell et al. [58] and hence there would not be enough reads for efficient junction detection. The authors state that the method introduced by Bryant et al. [7], which uses hashing as its alignment approach, needs a large amount of memory and computing power and as a result is not scalable for reads longer than 50 base pairs.

Huang et al. [23] state that SpliceMap, the algorithm presented by Au et al. [3], performs poorly when dealing with the reads that can be mapped to more than one location. Furthermore, they state that this approach is not efficient when the transcriptome is lowly expressed or the reads have sequencing errors. The authors state that the method developed by Wang et al. [63] has some inefficiencies while the sequencing depth is low, which leads to a reduced call rate. The call rate is the number of true positives divided by total number of junctions.

Huang et al. [23] present SOAPsplice, which finds the splice junctions in two steps. In the first step, it maps the reads onto the reference genome using the Burrows Wheeler Transformation for indexing. Then, SOAPSplice detects splice junction candidates based upon some criteria, which include following known splicing motifs and a maximum intron size of 50,000 bp. SOAPsplice applies two different filtering techniques to omit false positives. The first strategy is to check the paired-end information with the direction of the mate-pair reads and later discarding incompatible junctions. The other strategy is to filter out the junctions that have a missing segment between two sub-reads that have been mapped to the reference genome.

Huang et al. [23] compared their method with the algorithms developed by Trapnell et al. [58], Wang et al. [63], and Au et al. [3] on two 50 and 150 bp simulated datasets and two 51 and 130 bp real datasets.

The authors claim that based on the results of the simulated datasets for both 50 and 150 bp length reads, their method had the highest call rate while it kept the false positive rate at its lowest compared to other approaches. For the real dataset with 51 bp reads, Huang et al. [23] claim that SOAPsplice detects more novel junctions than TopHat by Trapnell et al. [58] and its results are comparable to the method designed by Au et al. [3] on both novel and known junctions. According to the authors' claim, SOAPsplice found more splice junctions than the other compared methods, and 97.24% of detected junctions were reported by more than one method. Huang et al. [23] claim that although their method found fewer novel junctions than the methods by Au et al. [3] and Wang et al. [63], but the percentage of junctions that are reported by more than one method for SOAPsplice (85.34%) is significantly higher than those of the other algorithms (TopHat: 67.73%, SpliceMap: 63.24%, MapSplice: 77.54%).

Huang et al. [23] claim that their method is more efficient for detecting novel splice junctions as it outperforms all other algorithms with various read lengths and read depths, especially when sequencing depth is lowest. This is very important considering that new junctions are usually found in low abundance parts of the transcript. The authors claim that their method is able to detect more genuine splice junctions than the compared methods.

As described by Zhang et al. [70], RNA-Seq can be used for "cellular phenotyping" and to help establish the etiology of diseases characterized by abnormal splicing patterns. Recent studies have revealed that variations in splicing patterns are associated with Alzheimer's

Year	Author	Package Name	Notes
2009	Trapnell et al.	TopHat	Been cited by all papers, that reviewed in this study, which published after it. The most cited paper in overall. The package is being updated very often and source code is freely available.
2010	Wang et al.	MapSplice	Cited by Huang et al. [23], and Zhang et al. [70] The package is being updated regularly and source code is freely available.
2011	Huang et al.	SOAPsplice	The package is being updated regularly and the package is freely available.
2012	Zhang et al.	PASSion	The latest published package reviewed in this study. Source code is freely available.

Table 3.3: Availability and update frequency of the software packages mentioned in this section.

and other complex diseases. In RNA-Seq, the exact nature of splicing events is buried in the reads that span exon-exon boundaries. The accurate and efficient mapping of these reads to the reference genome is a major challenge.

Zhang et al. [70] refer to previous works by Trapnell et al. [58], Dimon et al. [14], and Wang et al. [63]. The authors claim that the methods developed by these authors do not have the ability to detect junctions without known splicing motifs. Zhang et al. [70] state that both HMMSplicer by Dimon et al. [14] and MapSplice by Wang et al. [63] potentially work better for long reads than for short reads and they are less accurate on highly abundant transcripts. They also claim that neither of these two methods exploit the paired information in their algorithms.

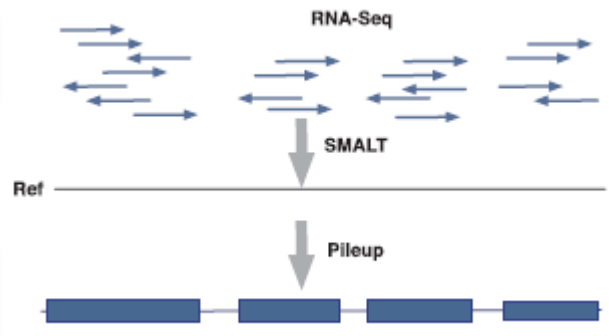

Figure 3.5: First step in PASSion pipeline, including mapping the reads using SMALT and creating exon islands [70]. Courtesy of Oxford Journals.

PASSion finds the splice junctions in five stages including the initial mapping, building exon islands, high-resolution remapping, filtering and detection of canonical and non-canonical junctions. As shown in Figure 3.5, exon islands are built by piling up the mapped reads after initial mapping by a fast aligner. Pairs of one exonic read and one unmapped read are used as the basis of junction identification. These pairs are remapped, using the pattern growth algorithm, to the reference genome and a splice junction is reported if the unique substrings from both ends can reconstruct the original split read and has a sufficiently large number of supportive reads. Figure 3.6 shows this process in detail.

Initial mapping and high-resolution remapping are the most time-consuming parts of the PASSion's pipeline. PASSion uses SMALT for the initial mapping of the reads to the reference genome. SMALT is an aligner that has been designed by the Sanger Institute [70], for aligning of DNA sequences to the reference genome. It only accepts sequence input data in FASTA or FASTQ format. SMALT uses a hash index of short sequences under 21 base pairs long.

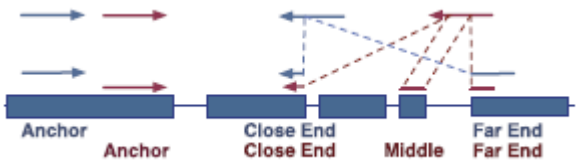

Figure 3.6: Using paired-end sequencing technology in splice junction discovery in PASSion. PASSion designates a mapped read with an unmapped pair as anchor, then uses the direction of the anchor to look for the other pair possibly mapped across a splice junction [70]. Courtesy of Oxford Journals.

Zhang et al. [70] analyzed their method on both simulated and real data. They compared the performance of PASSion with those of TopHat by Trapnell et al. [58], MapSplice by Wang et al. [63] and HMMSplicer by Dimon et al. [14] on these datasets. The authors claim that on simulated data, their method, alongside other three tested methods, can detect almost all the true junctions when coverage is $> 100\times$fold. They note that PASSion predicted 136,664 and 172,568 splicing events for the two real datasets, of which 84.1% and 80.3% are known junctions.

Zhang et al. [70] state that on the short read library of simulated data, the method by Trapnell et al. [58] showed the least sensitivity comparing to other methods, and on libraries with long reads, MapSplice by Wang et al. [63] detects the lowest number of junctions. The authors claim that in all simulated datasets, the true positive rate of PASSion has the quickest growth rate along with the read coverage and it is the most sensitive method overall. Zhang et al. [70] state that when the specificity of TopHat, MapSplice and HMMSplicer drops with the read coverage, PASSion's specificity remains high with specificities of more than 97%.

The authors claim that for real datasets, in general, PASSion displays a balanced performance with both a high number of predictions and high confirmed ratios. They state that

the pattern growth algorithm, which is used in their approach, has not been taken advantage of in RNA-Seq analysis before. Zhang et al. [70] note that PASSion can detect junctions with unknown motifs, which other three methods were unable to do so. Zhang et al. [70] state that their method had missed some rare cross-chromosome splicing events, because it has been assumed that two reads map to the same chromosome. They suggested working to resolve this issue in the future.

3.3 Conclusion

The work of Trapnell et al. [58] has been cited by 8 out of 10 papers that had been included in our review of splice junction detection methods. This means that except the work of De Bona et al. [13] that has been published prior to their work, all subsequent works on this subject referred to it. Furthermore, the method developed by Trapnell et al. [58], was used by all other methods as a basis for evaluating the performance of their own work. Table 3.4 lists the works studied in this chapter, and reviews their major contributions.

De Bona et al. [13] state that their method can be extended to exploit homo-polymer errors, which is available for Roche's 454 sequencing platform. Trapnell et al. [58] suggest that using paired-end reads will drastically reduce the number of false positives in TopHat, and also improves its performance.

Au et al. [3], Lou et al. [35], and Zhang et al. [70] developed their methods to exploit paired-end read information. Huang et al. [23] mention that in the future, their method could be optimized to run faster and consume less memory. We observed that SOAPSplice, the method presented by Huang et al. [23], has been updated after publishing the paper to reduce the amount of memory usage while generating the output.

Based on the studies on advantages and disadvantageous of various methods on splice

junctions discovery over each other, we chose to apply PASSion designed by Zhang et al. [70] in our study. The advantage of PASSion is that it had been originally designed to exploit paired-end information which is used in its mapping algorithm. Also the work by Zhang et al. [70] was the latest method developed on this topic studied in our review. Therefore, they had the chance to compare their results against previously developed methods. PASSion uses only known splicing motifs in the last step in its pipeline to finalize the breakpoint of a junction. As PASSion does not use known motifs in detecting junctions, it can detect junctions with unknown motifs.

Overall, PASSion showed a very high rate of accuracy in both high and low abundant transcripts [70]. The only downside of using PASSion is that under the same conditions, it consumes between two to four times more memory than the methods developed by Trapnell et al. [58], Dimon et al. [14], and Wang et al. [63]. Also, PASSion is the second slowest method among other methods in terms of CPU time.

Year	Author	Title of Paper	Major Contribution
2008	De Bona et al.	Optimal spliced alignments of short sequence reads.	**QPALMA**, one of the first works to address splice junction finding on RNA-Seq data. Use of SVM to find splice junctions.
2009	Trapnell et al.	**TopHat**: discovering splice junctions with RNA-Seq.	Introduces the concept of anchor as a way. Presents the idea of generating exon coverage islands.
2010	Au et al.	Detection of splice junctions from paired-end RNA-seq data by **SpliceMap**.	Designed to use information of paired-end reads. First method to use half-read mapping. Use of hash table for mapping.
2010	Ameur et al.	Global and unbiased detection of splice junctions from RNA-seq data.	**SplitSeek**, splits reads to two fragments and map them independently as anchors.
2010	Bryant et al.	**Supersplat**– spliced RNA-seq alignment.	Employs empirical RNA-Seq data for splice junction detection.
2010	Wang et al.	**MapSplice**: accurate mapping of RNA-seq reads for splice junction discovery.	Defining minimum anchor length as a filtering strategy. Comprehensive experiments on effect of various criteria including noise.
2010	Dimon et al. [14]	**HMMSplicer**: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data.	Employs hidden Markov model to determine the exon boundaries.
2011	Huang et al.	**SOAPsplice**: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data.	Claims to achieve a better performance than other major methods using more Memory and more computing power.
2011	Lou et al.	Detection of splicing events and multiread locations from RNA-seq data based on a geometric-tail (GT) distribution of intron length	Incorporates MLE method to align paired-end reads into reference genome. Introduces geometric-tail distribution for intron lengths.
2012	Zhang et al.	**PASSion**: A Pattern Growth Algorithm Based Pipeline for Splice Junction Detection in Paired-end RNA-Seq Data.	Introduces Pattern Growth algorithm to remap the reads. The ability to identify junctions with unknown splicing motifs.

Table 3.4: Splice junction discovery tools at a glance.

Part II

Methods

Chapter 4

Methods

In this chapter, we discuss the proposed methodology of our approach to the problem of finding biomarkers in detail. As Figure 4.1 illustrates an overview of the pipeline of the proposed method, we start by describing the details of the dataset that we have used and the pre-processing that enable for splice junction detection. As PASSion's algorithm has been discussed in the previous chapter, here we describe its parameters and details of operation for our study. Following the way, we come across the specifics of our algorithm used for filtering splice junctions, and selecting and proposing them as biomarkers and classification features. At last, we introduce SVM as our machine learning method of choice for the classification of the samples belonging to cancer and normal classes. Also, k-fold cross-validation is incorporated to validate the accuracy of our predictions.

4.1 Dataset

We have used a dataset consisting of raw RNA-Seq data belonging to 20 samples belonging to patients diagnosed with prostate adenocarcinoma and 10 matched benign prostate tissue

samples as our control population. None of the patients had received any preoperative therapy prior to radical prostatectomy. This dataset is publicly available as a GEO dataset with the Accession number GSE22260 [27].

The dataset contains more than 667 million paired-end RNA-Seq reads that have been acquired using the Illumina Genome Analyzer II platform. It includes 30 files in SRA format for 30 different samples. The dataset consists of short reads with of length 36 base pairs for both forward and reverse strands. Also, the insert size for the prostate cancer dataset is 150 bp.

4.1.1 Input data format

FASTA is one of the most well-known file formats used to represent and store nucleotide sequences, in which they are depicted by a sequence of characters. This text-based format is one of the formats that major databases such as NCBI accept as the input method to query their databases. Each read sequence in FASTA format contains a line of sequence description followed by the sequence itself. Different databases use their own template for the sequence description line to specify the format based on their needs.

FASTQ format is the successor of FASTA format that completes it by including the quality information for each read in the file. FASTQ has become the *de facto* standard for storing high-throughput sequencing technologies.

Most of the well-known software packages, as well as all packages that we use in this study, are only compatible with FASTA/FASTQ format. For this reason, we used SRA-Toolkit, developed by NCBI, to convert our dataset in SRA format to FASTQ format [41]. As SRA format stores both strands for reads in a single file, we needed to split each SRA file belonging to a sample into two separate FASTQ files to account for paired-end input

format of PASSion. We used “split-3” as a parameter in SRA-Toolkit in order to obtain the FASTQ files in paired-end format.

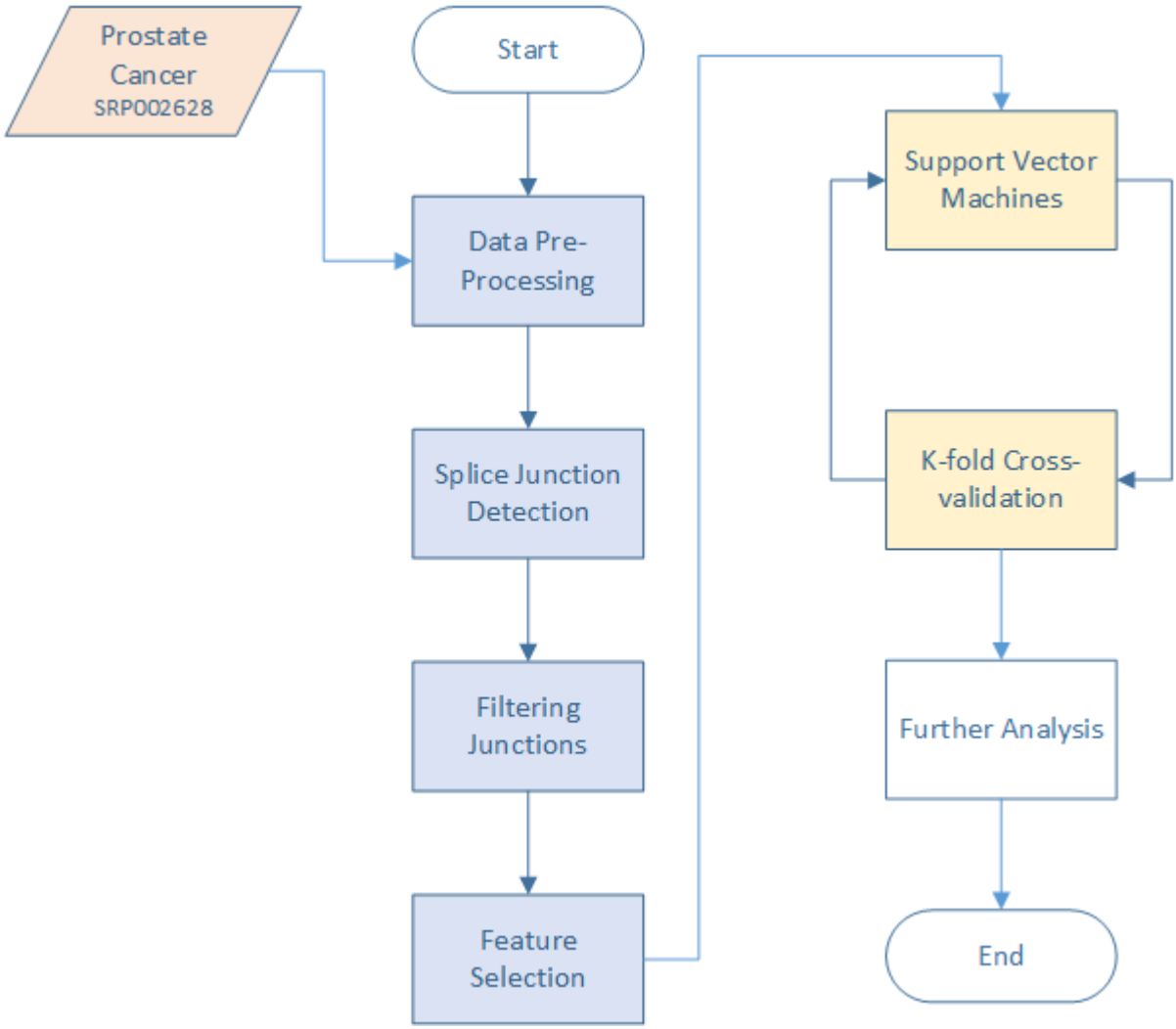

Figure 4.1: Pipeline for proposed model of this study.

4.2 Splice Junction Detection

The splice junction detection module is shown in details in Figure 4.2. In the following, we describe each part of this pipeline separately.

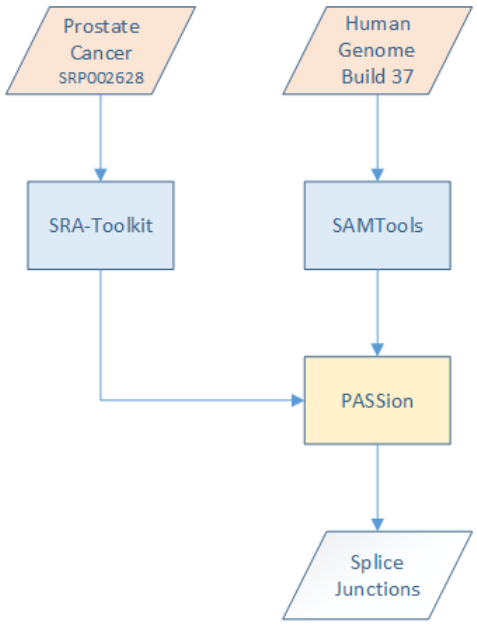

Figure 4.2: Splice junction detection pipeline.

4.2.1 Reference Genome

In order to successfully run PASSion, SMALT and SAMTools packages should be installed on the system as PASSion employs them in its pipeline. SMALT is a fast read aligner developed by the Wellcome Trust Sanger Institute which utilizes a hash index of short words in its algorithm. SMALT accepts reads and the reference genome in FASTA or FASTQ format. Prior to running PASSion, chromosome IDs used in the reference genome should be indexed separately by SAMTools. The resulting index file is used by PASSion as

an input. We used the latest version of the Human Genome, Build 37 (GRCh37.p10) from the Genome Reference Consortium [11], as our reference genome which acted as an input for PASSion and SAMTools. All mentioned software packages have been designed solely for Linux, and are publicly available.

4.2.2 PASSion

PASSion accepts five required and multiple optional arguments as input parameters. The required parameters are insert size, the paths to the two input read files, the reference sequence, the reference sequence index by SMALT. All optional arguments has been set to their default values recommended by Zhang et al. [70]. One of the important parameters, the cut-off limit, which is described in the following, has been set to 0.1. This parameter implies that any junction where its cut-off score falls short of the this limit will be discarded.

$$cut-off_{junction} = \frac{number\ of\ support\ reads}{coverage\ of\ higher\ expressed\ flanking\ exon} \tag{4.1}$$

Other important parameters include maximum number of SNPs allowed that is set to two. Minimum intron size has been set to 20, and sequence error rate is fixed at 0.05.

Detail output file

PASSion includes details about the mapping of the split reads across each exon-exon junctions in a file called Junctions.detail. A sample Detail file is depicted in Figure 4.3.

```
##########################################################################################################
1544 D 615 ChrID gi|224384768|gb|CM000663.1| BP 21807470 21808086 BP_range 21807470 21808091
Supports 1 + 0 - 1 S1 2 S2 18.123 LL 16 RL 20 SUM_MS 36 NumSupSamples 1 - 1
ACCAAGAAGAGGAAGAAGACCAAGGCCCACCATGCCcaggtaagt<595>ttctacttttCCAGGCTCAGC
                    CAAGGCCCACCATGCC                          CCAGGCTCAGCAGAGAGCTG           -21808102
36 -@SRR057629.2921017_1
```

Figure 4.3: Sample Junction.detail PASSion output file.

BED output file

PASSion stores the found splice junctions using the BED format. The BED format supplies the means to store data for an annotation track as standardized by the UCSC genome browser [6]. This format requires three fields as obligatory and nine fields as optional. The required fields store chromosome name, the starting position of the desired feature in chromosome, and ending position of that feature. Other fields could include information such as line name, score, strand direction, RGB value, block start and end, and block count. PASSion uses mainly chromosome start and end field to store break point ranges for each junction. According to Zhang et al. [70] starting and ending position of a junction can be calculated using the following equations:

$$Junction_{start\ position} = chromosome\ start\ position + block\ start \tag{4.2}$$

$$Junction_{end\ position} = chromosome\ end\ position - block\ end\ + 1 \tag{4.3}$$

```
gi|224384768|gb|CM000663.1| 21807454 21808105 JUNC_0 1 . 21807470 21808091 255,0,0 2 16,20 0,631
gi|224384768|gb|CM000663.1| 94953327 94953469 JUNC_1 3 . 94953345 94953449 255,0,0 2 18,24 0,118
gi|224384768|gb|CM000663.1| 53543454 53544066 JUNC_2 6 . 53543471 53544038 255,0,0 2 17,31 0,581
```

Figure 4.4: Sample Junction.bed PASSion output file.

PASSion saves block coordinates as the 11th field of the BED file, in which block start and end coordinates are separated by a comma. As an example, for the first read of the sample BED file in Figure 4.4, junction start position can be calculated as $Junction_{start}$: $21807454 + 16 = 21807470$ and $Junction_{end}$: $21808105 - 20 + 1 = 21808086$.

These starting and ending positions match the positions for the same read in the corresponding detail file after the phrase "BP" (Break-point), which is shown in Figure 4.3. The number 1 after the phrase "Supports", in the detail file, indicates that there is only one read supporting this junction. This means that the expression level for this junction is only one. This measurement number can also be found after the junction ID in the bed file (Figure 4.4).

4.3 Filtering Junctions

As we introduced 30 different samples which were processed separately by PASSion in our system, it was conceivable that PASSion could not account for the factor of error set forth by differential analysis of the input data. Our dataset included 20 samples in the cancer class and 10 samples in the benign class in total. Due to this low number of samples and also the high probability of base pair errors in the mapping process, we needed to identify the same junctions across all different samples with high accuracy. That is necessary as a single base pair error introduced in mapping, in either the starting position or the ending position of a

junction, could jeopardize the process for that junction. We developed a method to filter out dubious junctions to improve the accuracy of our method as well as to decrease the error rate. Different parts of our approach for addressing this issue are illustrated in Figure 4.5. This problem can be modeled as a peak-finding problem in a 3-dimensional space.

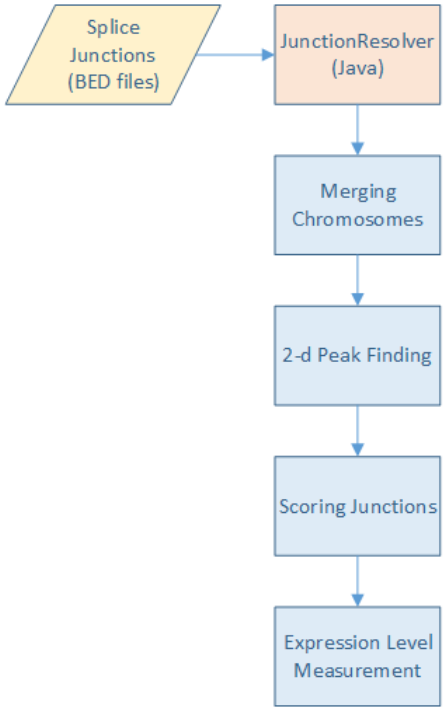

Figure 4.5: Modules for filtering junctions.

The solution to the problem of finding splice junctions as biomarker for cancer samples is based on alternative splicing, which means that in some samples that have been affected by cancer, mRNA has been spliced differently than the normal samples. We designed a fast and efficient algorithm to account for the inherent differences that have been introduced to the system by running peak finding separately for starting and ending positions.

Considering the fact that starting and ending positions of each junction are the boundaries between exons and introns, we expect in case of alternative splicing in each sample the same position would happen in other junctions but with different starting or ending positions. It is important to remember that studying alternative splicing only is possible because we are mapping our reads against the reference genome. Our method models the starting and the ending points as semi-independent entities to each other.

4.3.1 JunctionResolver

We developed a Java API called JunctionResolver to handle all processing involved in working with junctions. The Junction class contains the following fields: start, end, length, expLvl, chrom corresponding respectively to start and end position of a junction relative to the chromosome, length of the junction, number of reads supporting the junction, and chromosome ID of the junctions. Chromosome ID for chromosomes 1 to 22 are the same as their number, and we considered 23 and 24 for X and Y chromosomes for convenience.

A class called ChromJunc were designed to store a list of junctions belonging to each specific chromosome. Classes BedReader and DetailReader have been developed to read information from PASSion's output files. In order to differentiate between samples, we included the sample ID as a property in the BedReader class. We developed the ChromFinder class in order to find each junction in a list of junctions, which are stored as a ChromJunc object. A class named Comparator has been designed to compare junctions for two different ChromJunc objects and report the number of junctions that start or end at the same position.

We developed a solution to write data in Comma-separated values (CSV) format to store data for the next module of our pipeline. Figure 4.6 shows a sample of the output of the

JunctionResolver module in CSV format. This file includes the sample ID, the chromosome ID, starting position, ending position, junction's length, and supporting reads count, respectively for each junction.

```
29,1,21807470,21808086,616,1
29,1,94953345,94953446,101,3
29,1,53543471,53544036,565,6
```

Figure 4.6: Sample CSV output file.

4.3.2 Merging Chromosomes

As starting and ending positions of the junctions are relative to each chromosome, it was critical for our program to separate the junctions by chromosomes. As the first step, we used MATLAB to read CSV files belonging to each sample based on the sample ID in their file names, also CSV files included the sample ID as the first field for each junction. We separated the junctions for each chromosome in a different data structure. After this step, we would have a table with the size of $number\ of\ chromsomes \times number\ of\ samples$ cells. This table in our case included 24×30 cells. This table will be the input for our peak finding module as described in Algorithm 1. In the next step, we join all junctions for each chromosome in a single cell. The results will be a table of $1 \times number\ of\ chromosomes$ cells.

4.3.3 2-D Peak finding

We designed a module in MATLAB to create a histogram based on either starting or ending position of junctions, and this module was utilized on all chromosomes. The last position of a starting junction on the first chromosome was $249,231,781$ in our dataset. To account for the amount of memory required to process the data for the whole genome in this scale, we used the sparse matrix data structure in MATLAB to handle the problems arising from working with these huge tables. The solution was to split the table into several smaller windows, transform the data into a full matrix for each of them, and then process the data in each window separately. Finally, we merge the results together in a new sparse matrix structure. The size of the window is a parameter for our module that has been set to $100,000$ for our study.

We also implemented a safety mechanism to make sure that no peak occurs at the boundaries of a window. In order to do so, in case of a non-zero value at vicinity of a window, we move the window until we reach an empty space with at least a length of 5. We have developed a module, that uses MATLAB to find the rough peaks, to find junctions along the whole chromosome using the sparse matrix as described.

We define a parameter called *margin* to be passed to this module as a minimum peak distance variable, which defines the minimum distance between two peaks. After peak finding process on starting positions finishes, if position a is found as a peak, we search for all junctions that have a start position in the neighbourhood of $(a-margin, a+margin)$ and we acquire a set of ending points for the selected starting points. As our peak finding module discards some starting points near peaks, we use our margin parameter to account for them when analyzing end points in the next step. Considering the minimum length of a junction in our dataset which is $19bp$, we selected $5bp$ as the vicinity for combining junctions. This

gives a safe margin near a quarter of the size of the minimum intron. We chose $margin = 2$ to cover a $5bp$ area of the genome for each junction position.

Data: Junctions table of size $samples \times chromosomes$

Result: Set of merged junctions per chromosome

$direction \leftarrow$ "$start$";

$margin \leftarrow 2$;

for $k \leftarrow 1$ **to** *Number of chromosomes* **do**

  $Positions(k) \leftarrow CombineSamples(k, direction)$;

  $Peaks(k) \leftarrow findPeaks(Positions(k), margin)$;

  foreach *start position i of the Peaks(k)* **do**

    for $j \leftarrow i - margin$ **to** $i + margin$ **do**

      $endPositions(j) \leftarrow End\ points\ for\ all\ junctions\ starting\ on\ j$;

    end

    $endPositions(i) \leftarrow findPeaks(endPositions(j), margin)$;

  end

end

Algorithm 1: 2-d peak finding algorithm.

Ending points

So far, we have unified the starting points for all junctions in each chromosome across all samples. As can be seen in Algorithm 1, these unified starting points have been found by running peak finding modules for the first time. We continue by considering each starting point separately i of $Peaks(k)$. For each starting point, only end points are deemed fit that have their starting position in the vicinity of our unified starting positions, and hence these points act as a mean to limit the searching space to find the local maxima. To obtain the

final junctions, we run our peak finding algorithm on the ending points. We also stored sample IDs for each ending point of a junction.

4.4 Selecting Junctions

4.4.1 Scoring Junctions

We propose a method to score each junction based on the number and class of samples which that junction occurs in them. Given the fact that we have twice the number of cancer samples compared to our benign population, our model considers a $+1$ score for each sample that belongs to the cancer class, and a -2 score for each normal sample. The scoring formula is defined in Equation 4.4. This scoring scheme guarantees that a junction that has occurred in all samples, which implies no significance in terms of differentiating between classes, would receive a score of zero.

$$Score_{junction} = (No.\ of\ Junctions_{Normal} \times -2) + (No.\ of\ Junctions_{Cancer} \times 1) \quad (4.4)$$

4.4.2 Thresholding

We can limit the number of junctions reported by our filtering mechanism based on a defined minimum score. As high scoring junctions have occurred more frequently in only one class of our samples, they are expected to be more significant as features for a classification scheme.

4.4.3 Expression Level Measurement

The number of reads supporting each junction could act as a measure on the importance and impact of a particular junction across different samples. We exploit this information by creating a table of expression levels for each junction across all samples, which has scored higher than our score thresholds. We use this table as the feature set for our classification. Figure 4.7 shows a sample of this table containing expression level measures on three lines for three different junctions. The first 20 numbers on each line correspond to the expression measure for samples belonging to the cancer group and the remaining 10 numbers are expression measures for the samples in the benign group. A value of 0 for any sample implies that the specific junction has not been found for him, and any value greater than 0 shows the supporting number of reads for that junction.

```
0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,2,0,1,0,2
3,1,19,4,37,8,3,9,14,16,0,27,23,10,57,7,53,30,1,16,0,0,0,0,2,0,10,0,0,0
1,2,1,1,0,0,1,2,0,3,0,1,4,2,1,1,2,1,0,2,0,0,0,0,0,0,0,0,0,0
```

Figure 4.7: Sample expression level output table.

4.5 Classification

4.5.1 Support Vector Machines

Support vector machines are supervised learning methods, based on statistical learning theory, which are designed for classification and pattern recognition. SVM works by estimating a function called linear discriminant function that models the problem at hand [12; 62].

The basic SVM is a two-class linear classifier that is based on a linear discriminant function [4]. SVM could also be modeled as a non-linear classifier with the use of different kernel functions. We examine both linear SVM and non-linear SVM in our approach. We have tried various functions such as polynomial of degrees 2 and 3, radial basis, and sigmoid function as kernels.

In essence, SVM maps the input samples to a higher dimension feature space, and tries to find a hyperplane that separates the classes with the largest margin possible in the new space. In case that the problem is not linearly separable, SVM, based on the idea of the soft-margin, chooses a plane that separates the samples as clearly as possible.

In the present study, we used Weka with libSVM as well as libSVM implementation in MATLAB for our SVM implementation. Both of these software packages are freely available at their web sites [10; 21].

4.5.2 *K*-fold Cross-validation

In k-fold Cross-validation, the dataset is divided into k equal subsets. Each time, one of the k subsets is chosen as the prediction set and the other $k-1$ sets are used as training sets. The average of the k accuracy rates is the cross-validation accuracy rate. The accuracy rate is calculated by dividing the number of samples that has been classified correctly by the total number of samples. We have used k-fold cross-validation as a validation technique for our study in which we randomly divide our dataset into 10 subsets of equal sizes to perform classification.

Part III

Results and Discussion

Chapter 5

Results and Discussion

In this chapter, we present our results on steps that we take starting from finding splice junctions from raw reads, to filtering them, selecting junctions as biomarkers, and finally using them as features for the classification step.

5.1 Experimental Results

5.1.1 Reads

Our dataset contains $667,748,180$ reads in total, with an average of $24,597,860$ reads for cancer samples and $17,579,096$ for benign samples. This difference in number of reads could be explained by different levels of expression between cancer and benign samples. Figures 5.1 and 5.2 show the number of reads per sample for benign and cancer-diagnosed groups.

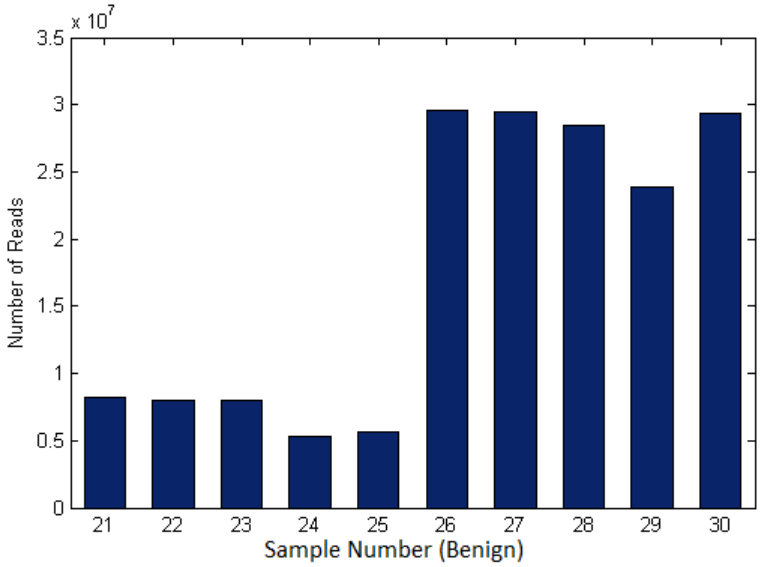

Figure 5.1: Number of reads among benign samples.

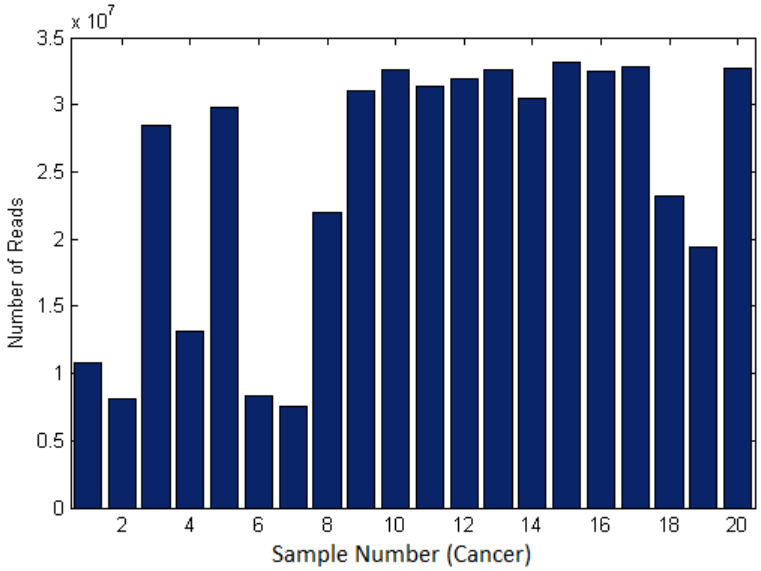

Figure 5.2: Number of reads among cancer samples.

5.1.2 Splice Junctions

We obtained a total of $3,272,686$ splice junctions from 30 samples using PASSion across all chromosomes. Total number of junctions per sample is shown in Figure 5.3. In this figure, the number of reads and number of junctions have been plotted against two different axes. The number of junctions is represented by bars, while the number of reads is shown by a dashed line. Figures 5.4 and 5.5 show the number of junctions found per sample for the benign and cancer samples respectively. The Pearson correlation coefficient of 0.97 between the number of reads and the number of junctions found per sample indicates a strong relationship between them.

Processing this data by PASSion took an estimated time of $2,200$ CPU hours using two $2.26GHz$ Intel Xeon CPUs server running Ubuntu 10.04. While it is possible to run as many threads as required using PASSion concurrently, due to the high memory consumption of each thread, that is near $9GB$, it was not possible for us to run more than 5 threads at a time considering our server's $48GB$ memory capacity.

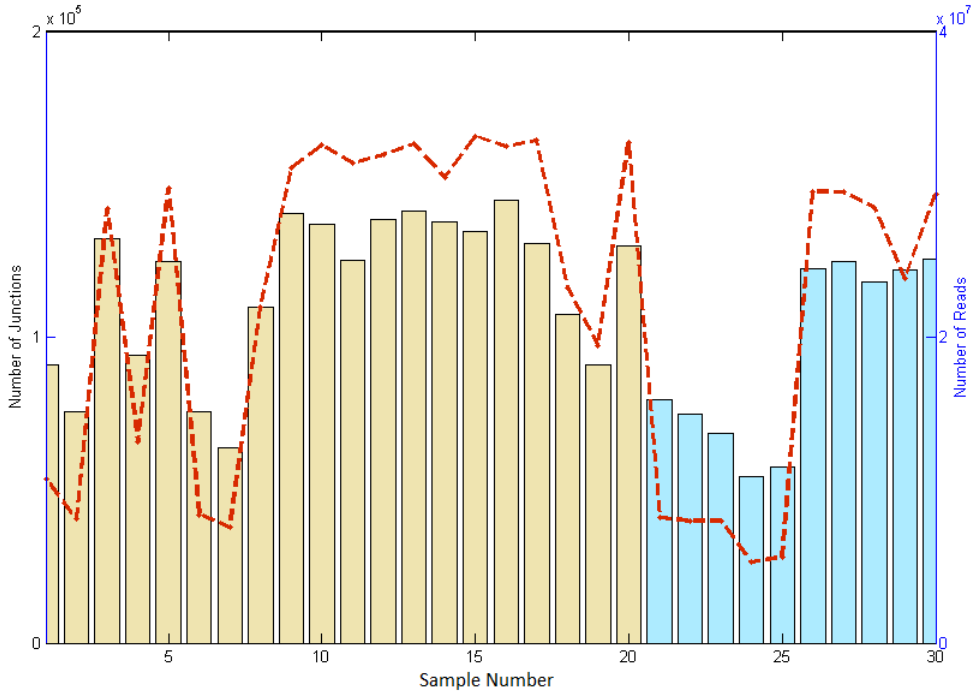

Figure 5.3: Number of reads and found junctions among all samples.

The average numbers of junctions found were 116, 267 and 94, 734 for cancer-diagnosed samples and benign samples respectively. The higher number of junctions for cancer-diagnosed samples indicates that higher levels of expression has lead to more splicing and hence higher number of junctions. This result regarding splice junctions, follows the same pattern as that observed by Kannan et al. [27], who studied this dataset previously, stating that paired chimeric reads are shown to happen more frequently in cancer samples than in benign samples.

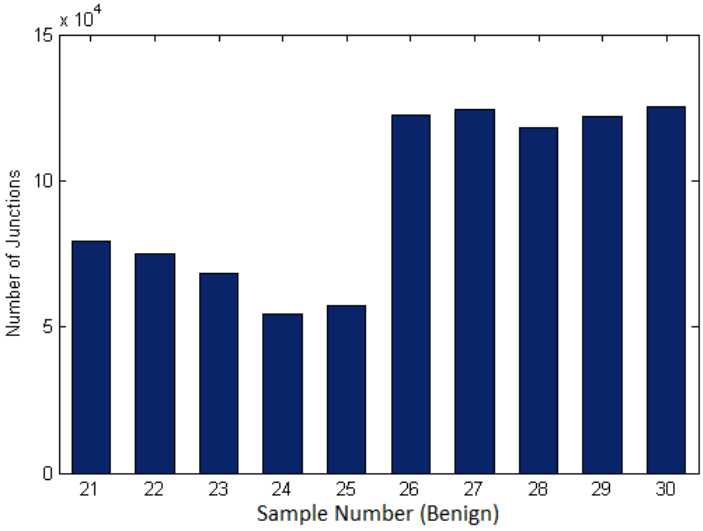

Figure 5.4: Number of junctions among benign samples.

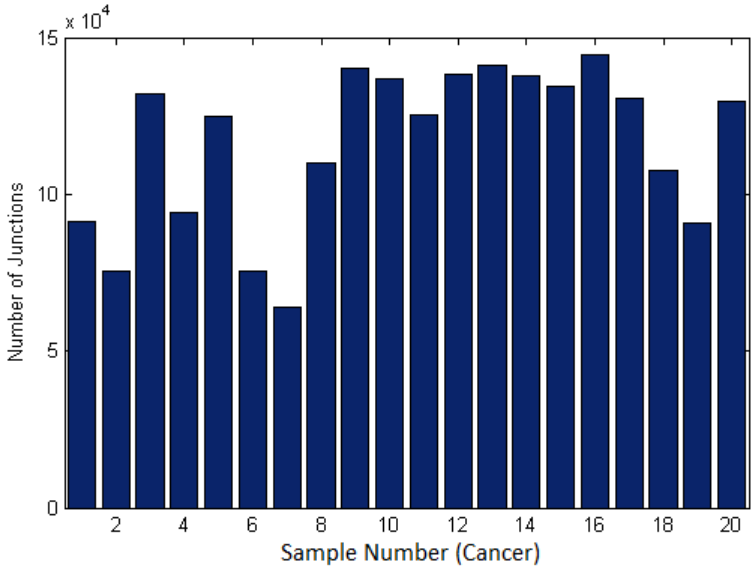

Figure 5.5: Number of junctions among cancer samples.

5.1.3 Junction Lengths

We calculated the length of the junctions across different samples and groups, and also across different chromosomes. As can be seen in Figure 5.6, there is no significant difference between the lengths of junctions across different chromosomes. The minimum length of junctions for all chromosomes is between 19 to 21 bp and the maximum length falls between $391,653$ to $409,637$ bp. However, after analyzing the average lengths of junctions across different samples, shown in Figure 5.7, it can be seen that the average length of junctions in samples is affected by the number of reads for each junction and subsequently number of found junctions. In this figure, samples belonging to the benign group are highlighted with a light color.

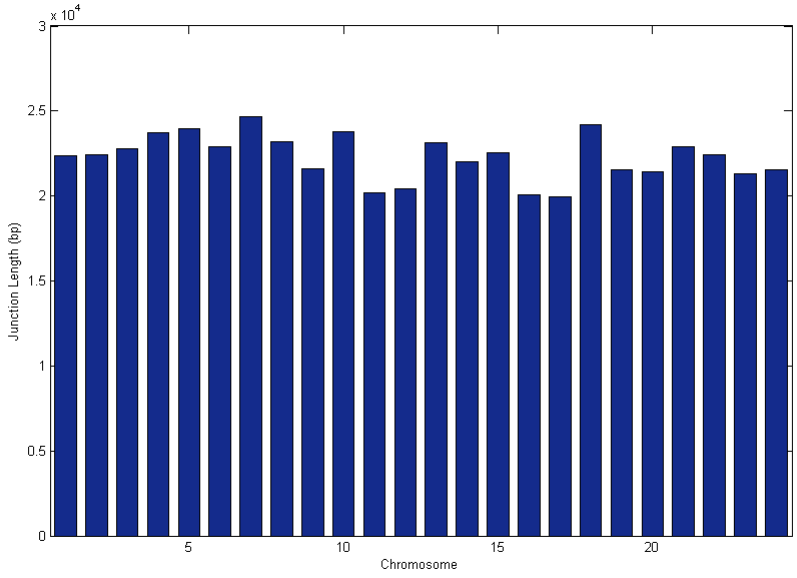

Figure 5.6: Average length of junctions across different chromosomes.

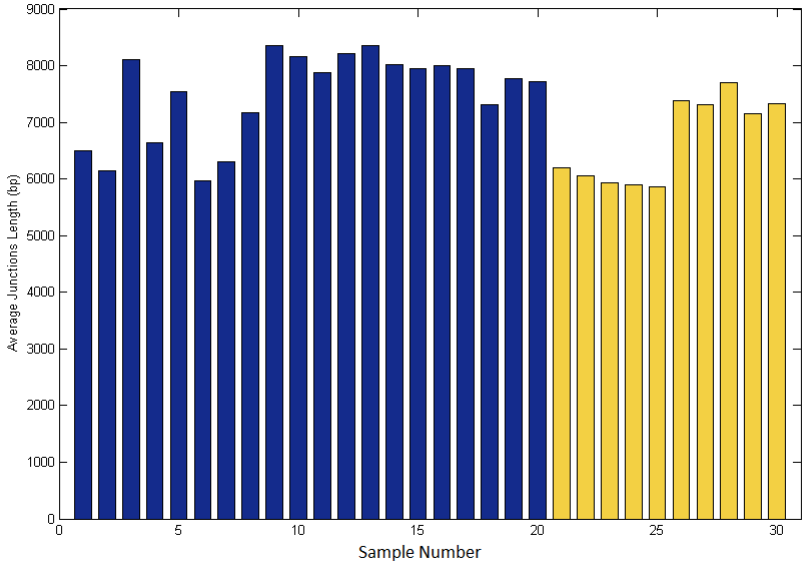

Figure 5.7: Average length of junctions across different samples.

5.1.4 Filtering

We merged and combined junctions from our population of 30 samples to find a set of distinct junctions to be used as features for prediction of prostate cancer. After this step, we applied our 2-D peak finding algorithm to remove dubious junctions. Figure 5.8 shows the number of junctions before and after the filtering process. As expected, most junctions have been found in Chromosome 1, which is the largest chromosome in the Human genome [52], and the least number of junctions has been observed in Y chromosome. Our filtering mechanism reduced the number of junctions by 6 to 8 percent by removing erroneous junctions by selecting 2 bp as our margin of error for the filtering process. Selecting a larger number for the margin leads to larger cuts in the number of junctions. Table A.1 contains full list of numbers of junctions before and after the filtering process for all chromosomes. We found

469,133 distinct junctions in total across all chromosomes using $2bp$ as margin.

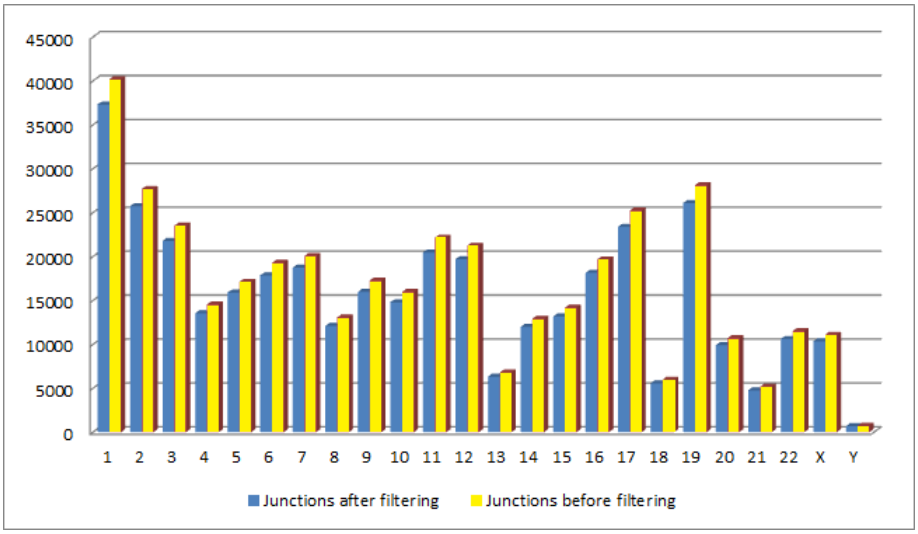

Figure 5.8: Number of junctions for each chromosome before and after filtering ($margin = 2bp$).

5.1.5 Scored Junctions

We calculated the frequency of junctions on different samples for each chromosome separately. Using our scoring scheme that has been previously described in Section 4.4.1, we score each junction using a number between -20 and 20, based on the number and class of the samples which the junction belongs to. We acquired 323,097 junctions with score greater than or equal to 1, and 146,036 junctions with a score lower than 1.

The histogram of junction scores for the first chromosome is shown in Figure 5.9. There are two main peaks seen in the histogram, these peaks happen at the scores of 1 and -2. This shows that the majority of junctions have been observed in only one sample. Each junction that has happened in a single cancer sample scores 1 and each junction from the benign group scores -2.

The local minima seen at score -1 is explained mostly by the junctions that have occurred at one sample belonging to cancer group and one sample from the normal group. Figures 5.9, 5.10, 5.11, show the histograms for chromosomes 1, 14 and Y. It can be observed from these figures that the shape of the histogram is almost identical across all these three chromosomes. The same pattern happens among all other chromosomes as well. A sharp peak is observed across all chromosomes at score 1.

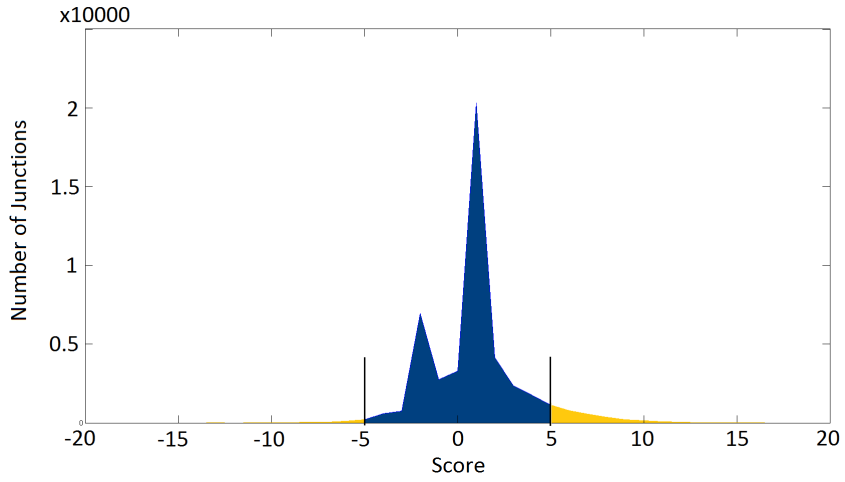

Figure 5.9: Histogram of junction scores for Chromosome 1.

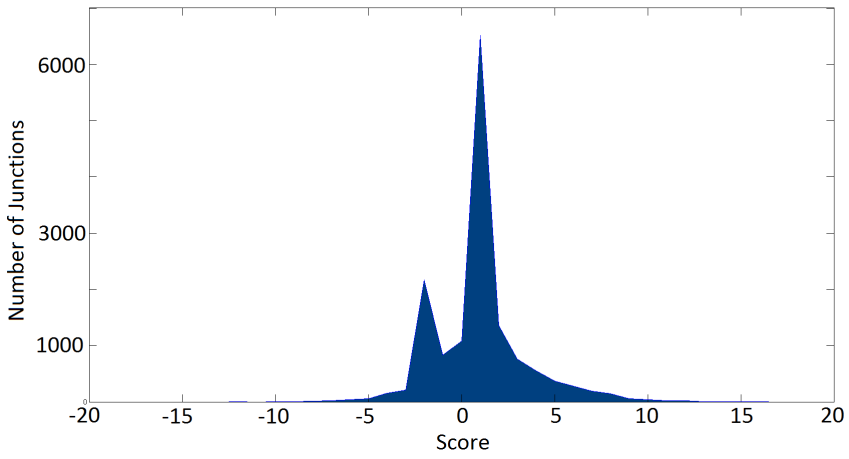

Figure 5.10: Histogram of junction scores for Chromosome 14.

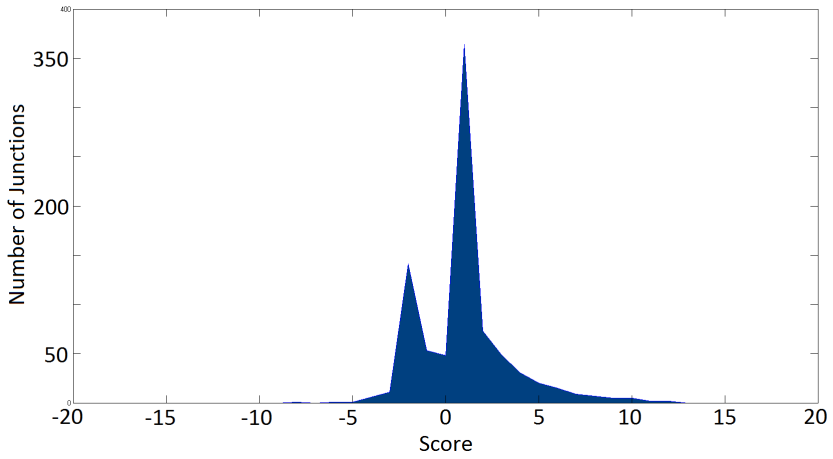

Figure 5.11: Histogram of junction scores for Chromosome Y.

5.1.6 Junction Selection

In order to select the junctions that better differentiate cancer group versus the benign group, we designated an *upper bound* and a *lower bound* cutoff value for our feature selection. This means that only junctions with scores more than the upper bound or with scores less than the lower bound will be considered in our junction selection process, implying that junctions that have happened more frequently for each class of samples have a higher chance to be selected as our features. This process would make our classification meaningful and also feasible given the vast number of junctions found from the dataset.

We applied classification on our scored junction dataset using SVM in order to predict the class of the samples. We used a grid-search approach to find cross-validation accuracies using SVM with different kernel functions. The grid-search approach enabled us to try a wide range of different feature sets, based on different upper and lower bounds, for our classification. The included kernels are: linear, radial basis, sigmoid and polynomial of degrees 2 and 3.

A range of 5 to 20 and -13 to -5, for upper and lower bounds respectively, have been selected as our grid search range. Junctions with scores ranging from -5 to 5 have been discarded altogether due to their insignificance in prediction and because of being at large numbers. The lower bound limit stops at -13, as there was no junction with a score less than that value. Table 5.1 shows different number of junctions selected based on different upper and lower bounds.

	-13	-12	-11	-10	-9	-8	-7	-6	-5
5	23,354	23,357	23,362	23,367	23,407	23,472	23,691	24,117	25,135
6	14,816	14,819	14,824	14,829	14,869	14,934	15,153	15,579	16,597
7	9,150	9,153	9,158	9,163	9,203	9,268	9,487	9,913	10,931
8	5,394	5,397	5,402	5,407	5,447	5,512	5,731	6,157	7,175
9	3,039	3,042	3,047	3,052	3,092	3,157	3,376	3,802	4,820
10	1,640	1,643	1,648	1,653	1,693	1,758	1,977	2,403	3,421
11	828	831	836	841	881	946	1,165	1,591	2,609
12	409	412	417	422	462	527	746	1,172	2,190
13	172	175	180	185	225	290	509	935	1,953
14	79	82	87	92	132	197	416	842	1,860
15	34	37	42	47	87	152	371	797	1,815
16	12	15	20	25	65	130	349	775	1,793
17	4	7	12	17	57	122	341	767	1,785
18	1	4	9	14	54	119	338	764	1,782
19	1	4	9	14	54	119	338	764	1,782
20	1	4	9	14	54	119	338	764	1,782

Table 5.1: Number of junctions used as features in the classification based on the scores.

The junctions considered for grid search, and consequently being in the feature set, are shown with a lighter color in comparison to all junctions for Chromosome 1 in Figure 5.9.

In each run, a different set of features was used as the input for our classification module. As shown in Figure 5.12, we achieved the best results using linear SVM. Table 5.2 shows the acquired accuracy rates based on the upper and lower bounds of our algorithm. The accuracy score in a square ranging from pairs of $(-13, 10)$ to $(-10, 14)$, in addition to

other pairs close to the square such as $(15, -9)$ was 100%, which is shown by a light shade in the figure. Darker shades on the right side of the figure shows the areas that the accuracy falls under 80%. This trend is observable also in Figures 5.13 and 5.14, which represents our classification results using polynomial and RBF kernels.

We report our results with fixed SVM parameters with values of $\varepsilon = 0.001$, $\gamma = 1$, as they proved to be the best in our runs. 100% 10-fold cross validation accuracy was gained using linear SVM with as low as 12 junctions with all being positives scores, meaning that they were more frequent in cancer samples.

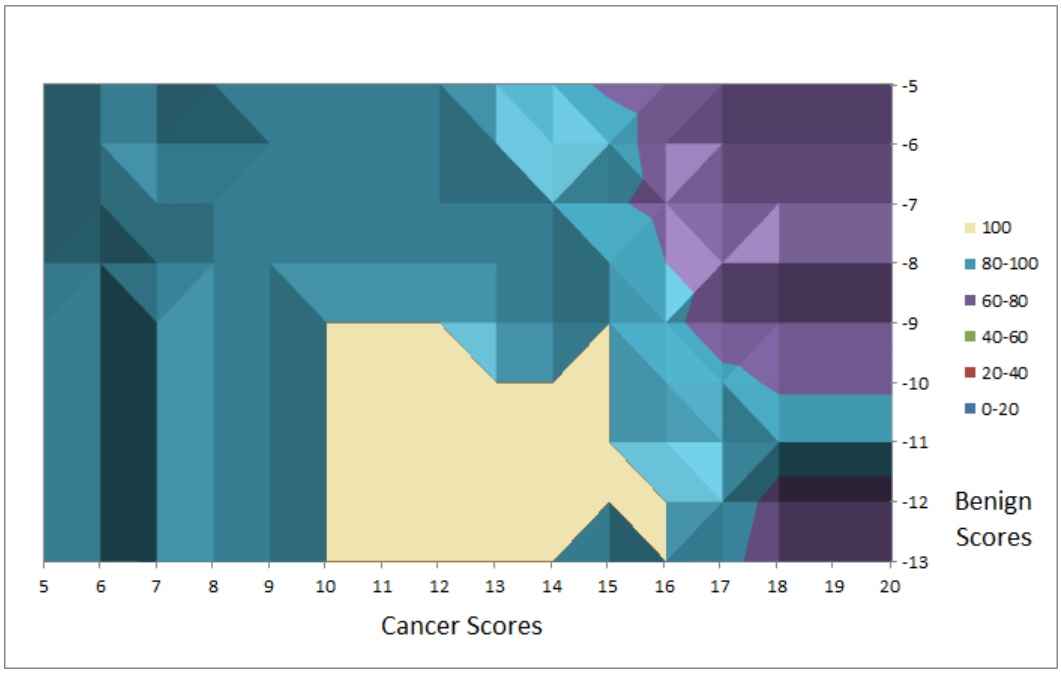

Figure 5.12: Accuracy of linear SVM classification.

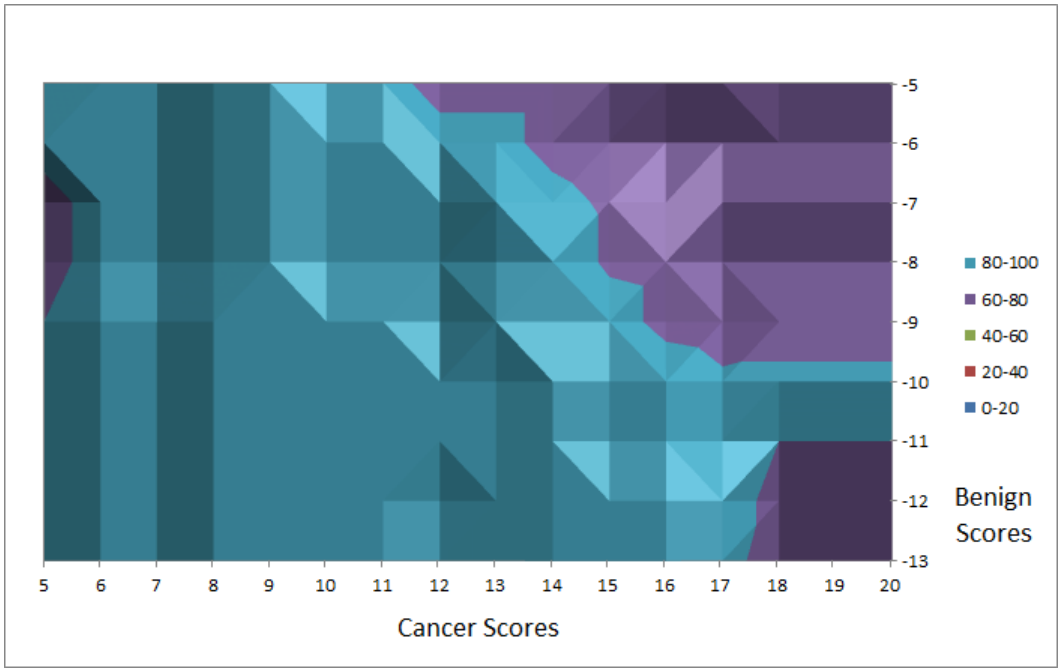

Figure 5.13: Accuracy of SVM with polynomial kernel (degree 2) classification.

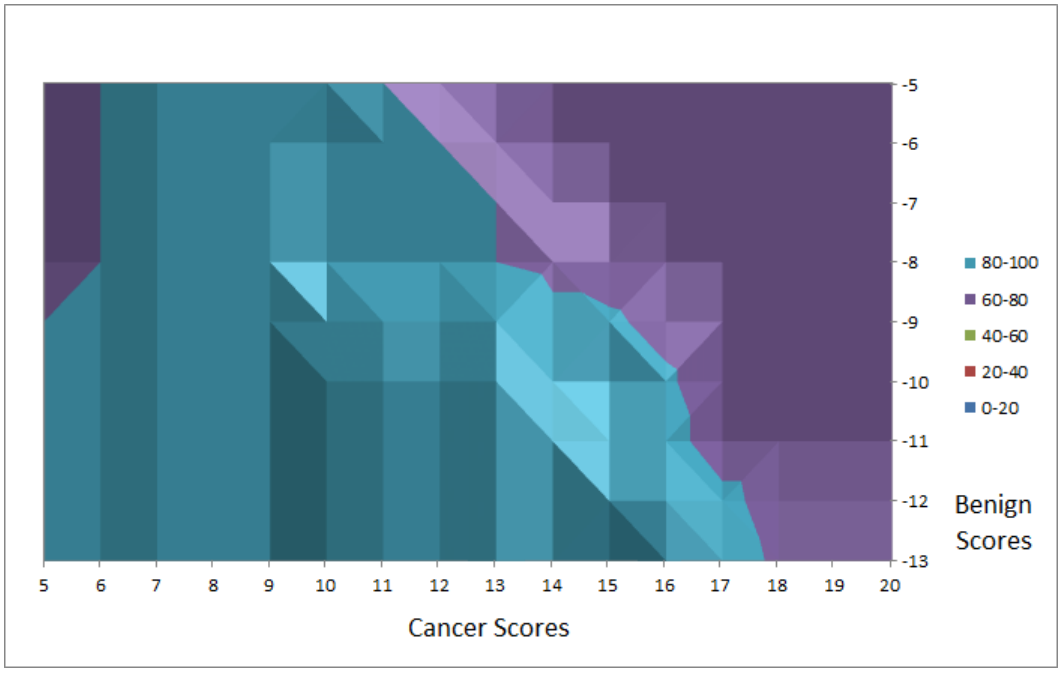

Figure 5.14: Accuracy of SVM with RBF kernel classification.

	-13	-12	-11	-10	-9	-8	-7	-6	-5
5	86.66	86.66	86.66	86.66	86.66	83.33	83.33	83.33	83.33
6	86.66	86.66	86.66	86.66	86.66	86.66	90.00	90.00	90.00
7	100.00	100.00	100.00	100.00	100.00	93.33	93.33	90.00	90.00
8	96.66	96.66	96.66	96.66	96.66	96.66	96.66	93.33	96.66
9	96.66	96.66	96.66	96.66	96.66	96.66	96.66	96.66	96.66
10	100.00	100.00	100.00	100.00	100.00	96.66	96.66	96.66	96.66
11	100.00	100.00	100.00	100.00	100.00	96.66	96.66	96.66	96.66
12	100.00	100.00	100.00	100.00	100.00	96.66	96.66	96.66	96.66
13	100.00	100.00	100.00	100.00	96.66	96.66	96.66	100.00	96.66
14	100.00	100.00	100.00	100.00	96.66	96.66	96.66	93.33	86.66
15	96.66	100.00	100.00	100.00	100.00	100.00	83.33	90.00	76.66
16	100.00	100.00	96.66	96.66	86.66	80.00	73.33	70.00	70.00
17	90.00	96.66	90.00	86.66	66.66	73.33	63.33	63.33	66.66
18	63.33	70.00	93.33	76.66	63.33	70.00	63.33	63.33	66.66
19	63.33	70.00	93.33	76.66	63.33	70.00	63.33	63.33	66.66
20	63.33	70.00	93.33	76.66	63.33	70.00	63.33	63.33	66.66

Table 5.2: Accuracy rates for linear SVM related to different scores.

5.1.7 Biological Analysis

To biologically analyze the results of our junction selection model, we focused on the 12 specific junctions belonging to the cancer-diagnosed group that we found using our scoring scheme and led us to 100% classification accuracy. We used BioMart [28] to find the corresponding gene for each of these junctions. In the next step, we used the Human Protein Atlas [60] to study previous annotations of each of these genes and their relationship prostate cancer. As shown in Table 5.3, cancer tissue staining was estimated at four different levels, including strong, moderate, weak, and negative. Based on this resource, these estimated numbers represent the percentage of samples that have been detected with prostate cancer antibody for each gene. The last column of this table, Normal Tissue Staining (NTS), represents the level of staining for that particular gene under normal conditions. We were then able to find cancer-staining information for 8 out of 12 studied genes. We

left the table cells empty for the genes that we could not find in the Human Protein Atlas or their impacts on various cancers were still under study.

Gene	Str.	Mod.	Weak	Neg.	Junction Start	Junction End	Score	Chr.	NTS
SRBD1	0	14	45	41	45812911	45826586	17	2	Medium
CRYBG3	0	18	9	73	97619328	97631010	18	3	Negative
ATP8A1	5	86	5	4	42454074	42457312	17	4	Medium
TRAPPC13					64957973	64960060	17	5	
FAM135A	0	46	25	29	71185250	71185364	17	6	Medium
POLR2J4					44056121	44058659	18	7	
PDE3B	0	81	19	0	14810786	14825488	17	11	Medium
XPOT	0	0	64	36	64811891	64812652	17	12	Weak
LEMD3	0	18	55	27	65634865	65637166	17	12	Weak
AC004696.2					56989748	57005530	18	19	
PLCB4					9351940	9352948	17	20	
CA5B	27	73	0	0	15792518	15793367	17	X	Weak

Table 5.3: Relationship of the genes containing selected features with prostate cancer.

5.2 Discussion

Based on the results of our study, we noticed that splice junctions happened more frequently in the cancer-diagnosed group compared to the normal group. Kannan et al. [27] also claim that based on their experiment, there are more chimeric RNAs in cancer than in benign samples. They state that this could be a sign of chimeric events as a result of cancer.

The fact that no junction with a score less than -13 has been found, implies the insignificance of junctions that only occur in samples belonging to benign tissues. The number of junctions observed with a score less than -10 is only 13. In other words, in more than $460,000$ junctions across all chromosomes, only 13 junctions have been found exclusively in more than half of the benign sample population. The corresponding number of junctions of the cancer group, with a score over 10, is $1,640$ junctions.

This observation matches the fact that we were able to reach 100% accuracy using only

junctions with positive scores, which we could not reach with negative scoring junctions. It also can be seen that in all versions of the SVM classification that we tried, depicted in Figures 5.12, 5.13, and 5.14, reducing the number of junctions with positive scores significantly reduces the accuracy of our classification. This effect is not significant for the opposite, meaning that accuracy of classification is not impacted heavily by reducing junctions from normal group in the feature set.

By observing the histograms for scored junctions across all chromosomes, we were unable to notice any significance for any chromosome. We also observed that linear classification leads to the best prediction results. Among other kernels, polynomial of degree 2 performed better than all other kernels including polynomial of degree 3. Also, sigmoid and radial basis function kernels performed the worst among all the classifiers. This could be due to overfitting in higher dimensions, indicating the linear nature of the problem [18].

We were able to spot a junction from gene CA5B in the X chromosome with a score of 17. According to the Human Protein Atlas project, this gene has 27% strong and 73% moderate staining property. This detection is significant as the aforementioned gene has a weak staining property for normal samples. This could hint at further studies towards splicing variants regarding this specific part of the gene. Also regarding the 8 genes that we found information regarding their relationship with prostate cancer, they were 42% stained moderately on average for prostate cancer. This fact supports our observation of higher expression levels in prostate cancer samples for higher scoring junctions in our study.

Part IV

Conclusion

Chapter 6

Conclusions

In this research, our goal was to exploit the paired-end information of RNA-Seq reads to extract biological meaning across a low population of samples. The sheer amount of data in our dataset added a new level of complexity to the RNA-seq experiments. Another challenge that we faced was differential detection of junctions across all samples in the population.

Although the effects of alternative splicing on prostate cancer has been previously studied [20; 40; 57; 9], and Ren et al. [47] studied prostate cancer in the Chinese population using alternative splicing on RNA-Seq data, we developed a novel model to apply machine learning methods on RNA-Seq dataset to select junctions as features and classified prostate cancer samples.

After studying current methods available for splice junction discovery, we selected PASSion to detect splice junctions for each individual sample. We designed and developed an algorithm to combine and merge these individual results and used an scoring scheme to select our junctions as biomarkers and consequently use them as features for our prediction module. We tried support vector machines with different kernels and on different sets of

features. A 100% 10-fold cross-validation accuracy has been achieved using linear SVM for different sets of junctions as features. Finally, we did research on the smallest set of junctions to compare with the previous findings regarding to prostate cancer on the genes that those junctions were belonging to. We found out that at least in one of the genes, there is an indication of significant relation to prostate cancer.

6.1 Contributions

- Design of a model for differential splice junction detection on prostate cancer data on large scale.
- Developing a filtering and scoring model as a feature selection mechanism leading to introducing junctions as biomarkers.
- Design of a machine learning based model for prediction of prostate cancer based on the selected features.

6.2 Future Work

Although we have used number of supporting reads for each junction as part of our features for the classification process, further analysis is needed in order to discover the probable effect of using expression level measurements in the feature selection process. Combining expression levels with our scoring algorithm might lead to more accurate results. Also combining different known feature selection algorithms could improve the results of our study.

Further studies regarding more accurate mapping of reads to the reference genome,

which will lead to more accurate splice junctions could lead to better accuracies in our study. Specifically, exploiting quality information of the reads stored in FASTQ format should enable PASSion to find more precise expression level measurements and more accurate junction positions.

Conducting pathway analysis could lead to a deeper biological insight into the result of the junction selection process, as well as the biological processes associated with prostate cancer.

Part V

Appendices

Appendix A

Supplemental Results

Number of chromosome	Junctions after the filtering	Junctions by PASSion
1	37309	40184
2	25718	27683
3	21779	23512
4	13537	14524
5	15896	17111
6	17841	19275
7	18718	20068
8	12092	13047
9	15967	17257
10	14789	15972
11	20453	22120
12	19674	21228
13	6315	6795
14	11970	12902
15	13165	14173
16	18150	19669
17	23382	25247
18	5543	5959
19	26088	28128
20	9891	10699
21	4763	5167
22	10605	11498
X	10330	11080
Y	666	710

Table A.1: Number of splice junctions for each chromosome before and after the filtering process (Margin = 2).

Appendix B

Guide for Running the Software Tools

B.1 Running PASSion

For setting up the respective modules, the following settings should be applied:

- SRA-Toolkit usage: `$fastq-dump --split-3 read.sra`
- SMALT usage: `$smalt index -k 13 -s 6 refindex reference.fa`
- SAMTools: Latest version of *samtools* installed form ubuntu packages
  - usage: `$samtools faidx reference.fa`
- PASSion ver 1.2.1 has been installed.
  - PASSion usage: `$passion.pl -s 150 -r read1.fq -f read2.fq -R reference.fa -I refindex -o passion_output`

B.2 Junction Dataset

- Dataset of raw junctions is available in BED and DETAIL format upon request.

B.3 JunctionResolver

- JRE 1.7 should be installed to run this package.
- *Readme* file is provided with the package as a guide for installation and runnning.

B.4 2-D peak finding and Junction Selection

- MATLAB installation is needed to run modules developed in MATLAB language.
- Full documentation is provided in *Readme* file.

Bibliography

[1] Bruce Alberts, Alexander Johnson, Julian Lewis, Martin Raff, Keith Roberts, and Peter Walter. *Molecular biology of the cell*. Garland, 4th edition, 2002. ISBN 0815332181.

[2] Adam Ameur, Anna Wetterbom, Lars Feuk, and Ulf Gyllensten. Global and unbiased detection of splice junctions from RNA-seq data. *Genome biology*, 11(3):R34, January 2010. ISSN 1465-6914. doi: 10.1186/gb-2010-11-3-r34.

[3] Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. *Nucleic acids research*, 38 (14):4570–8, August 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq211.

[4] Asa Ben-Hur and Jason Weston. A user's guide to support vector machines. *Methods in molecular biology (Clifton, N.J.)*, 609:223–39, January 2010. ISSN 1940-6029. doi: 10.1007/978-1-60327-241-4_13.

[5] Brigitta M N Brinkman. Splice variants as cancer biomarkers. *Clinical biochemistry*, 37(7):584–94, July 2004. ISSN 0009-9120. doi: 10.1016/j.clinbiochem.2004.05.015.

[6] UCSC Genome Browser. UCSC genome browser - FAQ: Data file formats, 2013. URL `http://genome.ucsc.edu/FAQ/FAQformat.html`.

[7] Douglas W Bryant, Rongkun Shen, Henry D Priest, Weng-Keen Wong, and Todd C Mockler. Supersplat–spliced RNA-seq alignment. *Bioinformatics (Oxford, England)*, 26(12):1500–5, June 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq206.

[8] Paul D Burns. TrueSight-ES: Machine learning approach to detecting splice junctions in RNA-Seq data. In *Plant and Animal Genome XXI Conference*. Plant and Animal Genome, 2013.

[9] Russ P Carstens, James V Eaton, Hannah R Krigman, Philip J Walther, Mariano A Garcia-Blanco, et al. Alternative splicing of fibroblast growth factor receptor 2 (fgf-r2) in human prostate cancer. *Oncogene*, 15(25):3059, 1997.

[10] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[11] Genome Reference Consortium. Genome reference consortium human build 37 patch release 10 (grch37.p10), 2012. URL `http://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.22/`.

[12] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[13] Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics (Oxford, England)*, 24(16):i174–80, August 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn300.

[14] Michelle T Dimon, Katherine Sorber, and Joseph L DeRisi. HMMSplicer: A tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq

data. *PloS one*, 5(11):e13875, January 2010. ISSN 1932-6203. doi: 10.1371/journal.pone.0013875.

[15] Huijuan Feng, Zhiyi Qin, and Xuegong Zhang. Opportunities and methods for studying alternative splicing in cancer with RNA-Seq. *Cancer letters*, November 2012. ISSN 1872-7980. doi: 10.1016/j.canlet.2012.11.010.

[16] Sergei a Filichkin, Henry D Priest, Scott a Givan, Rongkun Shen, Douglas W Bryant, Samuel E Fox, Weng-Keen Wong, and Todd C Mockler. Genome-wide mapping of alternative splicing in *Arabidopsis thaliana*. *Genome research*, 20(1):45–58, January 2010. ISSN 1549-5469. doi: 10.1101/gr.093302.109.

[17] Jessyka Fortin, Anne-Marie Moisan, Martine Dumont, Gilles Leblanc, Yvan Labrie, Francine Durocher, Paul Bessette, Peter Bridge, Jocelyne Chiquette, Rachel Laframboise, Jean Lépine, Bernard Lespérance, Roxanne Pichette, Marie Plante, Louise Provencher, Patricia Voyer, and Jacques Simard. A new alternative splice variant of BRCA1 containing an additional in-frame exon. *Biochimica et biophysica acta*, 1731 (1):57–65, October 2005. ISSN 0006-3002. doi: 10.1016/j.bbaexp.2005.08.011.

[18] Tobias Glasmachers and Christian Igel. Maximum likelihood model selection for 1-norm soft margin SVMs with multiple parameters. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1522–1528, 2010.

[19] Patrick Goymer. Transcriptomics: There's nothing abnormal about chimeric RNA. *Nature Reviews Genetics*, 9:734, 2008. ISSN 14710056. doi: 10.1038/nrg2459.

[20] Zhiyong Guo, Xi Yang, Feng Sun, Richeng Jiang, Douglas E Linn, Hege Chen, Hegang Chen, Xiangtian Kong, Jonathan Melamed, Clifford G Tepper, et al. A novel

androgen receptor splice variant is up-regulated during prostate cancer progression and promotes androgen depletion–resistant growth. *Cancer research*, 69(6):2305–2313, 2009.

[21] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, 11:10–18, 2009. ISSN 19310145. doi: 10.1145/1656274.1656278.

[22] Neil Hall. Advanced sequencing technologies and their wider impact in microbiology. *The Journal of experimental biology*, 210(Pt 9):1518–25, May 2007. ISSN 00220949. doi: 10.1242/jeb.001370.

[23] Songbo Huang, Jinbo Zhang, Ruiqiang Li, Wenqian Zhang, Zengquan He, Tak-Wah Lam, Zhiyu Peng, and Siu-Ming Yiu. SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Frontiers in Genetics*, 2(July):1–12, 2011. ISSN 1664-8021. doi: 10.3389/fgene.2011.00046.

[24] Clyde A Hutchison. DNA sequencing: bench to bedside and beyond. *Nucleic Acids Research*, 35:6227–6237, 2007. doi: 10.1093/nar/gkm688.

[25] Illumina Inc. Paired-end sequencing | achieve maximum coverage across the genome, 2010. URL `http://www.illumina.com/technology/paired_end_sequencing_assay.ilmn`.

[26] Illumina Inc. Estimating sequencing coverage, October 2011. URL `http://res.illumina.com/documents/products/technotes/technote_coverage_calculation.pdf`.

[27] Kalpana Kannan, Liguo Wang, Jianghua Wang, Michael M Ittmann, Wei Li, and Laising Yen. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. *Proceedings of the National Academy of Sciences of the United States of America*, 108(22):9172–7, May 2011. ISSN 1091-6490.

[28] Arek Kasprzyk. BioMart: driving a paradigm change in biological data management. *Database the journal of biological databases and curation*, 2011:bar049, 2011. ISSN 17580463. doi: 10.1093/database/bar049.

[29] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*, 10(3):R25, January 2009. ISSN 1465-6914. doi: 10.1186/gb-2009-10-3-r25.

[30] Heng Li and Richard Durbin. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25:1754–1760, 2009. doi: 10.1093/bioinformatics/btp324.

[31] Ruiqiang Li, Yingrui Li, Karsten Kristiansen, and Jun Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics (Oxford, England)*, 24(5):713–4, March 2008. ISSN 1367-4811. doi: 10.1093/bioinformatics/btn025.

[32] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25:1966–7, 2009. ISSN 13674811. doi: 10.1093/bioinformatics/btp336.

[33] Yang Li, Jeremy Chien, David I Smith, and Jian Ma. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, 27:1708–1710, 2011.

[34] Harvey Lodish, Arnold Berk, Chris A. Kaiser, Monty Krieger, Matthew P. Scott, Anthony Bretscher, Hidde Ploegh, and Paul Matsudaira. *Molecular Cell Biology (Lodish, Molecular Cell Biology)*. W. H. Freeman, 6th edition, June 2007. ISBN 0716776014.

[35] Shao-Ke Lou, Jing-Woei Li, Hao Qin, Aldrin Yim, Leung-Yau Lo, Bing Ni, Kwong-Sak Leung, Stephen Tsui, and Ting-Fung Chan. Detection of splicing events and multiread locations from RNA-seq data based on a geometric-tail (GT) distribution of intron length. *BMC Bioinformatics*, 12(Suppl 5):S2, 2011. ISSN 1471-2105. doi: 10.1186/1471-2105-12-S5-S2.

[36] Christopher A Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234):97–101, March 2009. ISSN 1476-4687. doi: 10.1038/nature07638.

[37] Christopher A Maher, Nallasivam Palanisamy, John C Brenner, Xuhong Cao, Shanker Kalyana-Sundaram, Shujun Luo, Irina Khrebtukova, Terrence R Barrette, Catherine Grasso, Jindan Yu, Robert J Lonigro, Gary Schroth, Chandan Kumar-Sinha, and Arul M Chinnaiyan. Chimeric transcript discovery by paired-end transcriptome sequencing. *PNAS*, 106:12353–12358, 2009.

[38] Andrew McPherson, Fereydoun Hormozdiari, Abdalnasser Zayed, Ryan Giuliany, Gavin Ha, Mark G F Sun, Malachi Griffith, Alireza Heravi Moussavi, Janine Senz, Nataliya Melnyk, Marina Pacheco, Marco A Marra, Martin Hirst, Torsten O Nielsen, S Cenk Sahinalp, David Huntsman, and Sohrab P Shah. deFuse: An Algorithm for

Gene Fusion Discovery in Tumor RNA-Seq Data. *PLoS Computational Biology*, 7: 16, 2011. doi: 10.1371/journal.pcbi.1001138.

[39] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature methods*, 5(7):621–8, July 2008. ISSN 1548-7105. doi: 10.1038/nmeth.1226.

[40] Goutham Narla, Analisa DiFeo, Helen L Reeves, Daniel J Schaid, Jennifer Hirshfeld, Eldad Hod, Amanda Katz, William B Isaacs, Scott Hebbring, Akira Komiya, et al. A germline dna polymorphism enhances alternative splicing of the KLF6 tumor suppressor gene and is associated with increased prostate cancer risk. *Cancer research*, 65(4):1213–1222, 2005.

[41] National Center for Biotechnology Information. SRA Handbook, 2010. URL `http://www.ncbi.nlm.nih.gov/books/NBK47537/`.

[42] Naoko Okumura, Hitomi Yoshida, Yasuko Kitagishi, Yuri Nishimura, and Satoru Matsuda. Alternative splicings on p53, BRCA1 and PTEN genes involved in breast cancer. *Biochemical and biophysical research communications*, 413(3):395–9, September 2011. ISSN 1090-2104. doi: 10.1016/j.bbrc.2011.08.098.

[43] Qun Pan, Ofer Shai, Leo J Lee, Brendan J Frey, and Benjamin J Blencowe. Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics*, 40:1413–5, 2008. ISSN 15461718. doi: 10.1038/ng.259.

[44] Dorothee Pflueger, Stéphane Terry, Andrea Sboner, Lukas Habegger, Raquel Esgueva, Pei-Chun Lin, Maria a Svensson, Naoki Kitabayashi, Benjamin J Moss, Theresa Y

MacDonald, Xuhong Cao, Terrence Barrette, Ashutosh K Tewari, Mark S Chee, Arul M Chinnaiyan, David S Rickman, Francesca Demichelis, Mark B Gerstein, and Mark a Rubin. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome research*, 21(1):56–67, January 2011. ISSN 1549-5469. doi: 10.1101/gr.110684.110.

[45] Dorothee Pflueger, Stéphane Terry, Andrea Sboner, Lukas Habegger, Raquel Esgueva, Pei-Chun Lin, Maria A Svensson, Naoki Kitabayashi, Benjamin J Moss, Theresa Y MacDonald, Xuhong Cao, Terrence Barrette, Ashutosh K Tewari, Mark S Chee, Arul M Chinnaiyan, David S Rickman, Francesca Demichelis, Mark B Gerstein, and Mark A Rubin. Discovery of non-ETS gene fusions in human prostate cancer using next-generation RNA sequencing. *Genome Research*, 21:56–67, 2011. doi: 10.1101/gr.110684.110.

[46] John R Prensner, Matthew K Iyer, O Alejandro Balbin, Saravana M Dhanasekaran, Qi Cao, J Chad Brenner, Bharathi Laxman, Irfan A Asangani, Catherine S Grasso, Hal D Kominsky, Xuhong Cao, Xiaojun Jing, Xiaoju Wang, Javed Siddiqui, John T Wei, Daniel Robinson, Hari K Iyer, Nallasivam Palanisamy, Christopher A Maher, and Arul M Chinnaiyan. Transcriptome sequencing across a prostate cancer cohort identifies PCAT-1, an unannotated lincRNA implicated in disease progression. *Nature Biotechnology*, 29:742–749, 2011. doi: 10.1038/nbt.1914.

[47] Shancheng Ren, Zhiyu Peng, Jian-Hua Mao, Yongwei Yu, Changjun Yin, Xin Gao, Zilian Cui, Jibin Zhang, Kang Yi, Weidong Xu, et al. RNA-Seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated

long noncoding rnas and aberrant alternative splicings. *Cell research*, 22(5):806–821, 2012.

[48] Iman Rezaeian, Yifeng Li, Martin Crozier, Eran Andrechek, Alioune Ngom, Luis Rueda, and Lisa Porter. Identifying informative genes for prediction of breast cancer subtypes. In Alioune Ngom, Enrico Formenti, Jin-Kao Hao, Xing-Ming Zhao, and Twan Laarhoven, editors, *Pattern Recognition in Bioinformatics*, volume 7986 of *Lecture Notes in Computer Science*, pages 138–148. Springer Berlin Heidelberg, 2013. ISBN 978-3-642-39158-3.

[49] David S Rickman, Dorothee Pflueger, Benjamin Moss, Vanessa E VanDoren, Chen X Chen, Alexandre de la Taille, Rainer Kuefer, Ashutosh K Tewari, Sunita R Setlur, Francesca Demichelis, and Mark A Rubin. SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. *Cancer research*, 69(7):2734–8, April 2009. ISSN 1538-7445. doi: 10.1158/0008-5472.CAN-08-4926. URL `http://www.ncbi.nlm.nih.gov/pubmed/19293179`.

[50] Paul Ryvkin, Yuk Yee Leung, Li-San Wang, and Brian D Gregory. Invited: Multiclass rna function classification using next-generation sequencing. In *Computational Advances in Bio and Medical Sciences (ICCABS), 2011 IEEE 1st International Conference on*, pages 10–10. IEEE, 2011.

[51] Anirban Sahu, Matthew K Iyer, and Arul M Chinnaiyan. Insights into Chinese prostate cancer with RNA-seq. *Cell research*, 22(5):786–8, May 2012. ISSN 1748-7838. doi: 10.1038/cr.2012.50.

[52] Meena Kishore Sakharkar, Vincent T K Chow, and Pandjassarame Kangueane. Dis-

tributions of exons and introns in the human genome. *In silico biology*, 4(4):387–93, January 2004. ISSN 1386-6338.

[53] Bernhard Schölkopf, K Tsuda, and J P Vert. *Kernel Methods in Computational Biology*, volume 11. 2004.

[54] Rolf I Skotheim and Matthias Nees. Alternative splicing in cancer: noise, functional, or systematic? *The international journal of biochemistry & cell biology*, 39(7-8): 1432–49, January 2007. ISSN 1357-2725. doi: 10.1016/j.biocel.2007.02.016.

[55] American Cancer Society. Cancer facts & figures 2013. Technical report, American Cancer Society, Atlanta, GA, 2013.

[56] Kyle Strimbu and JA Tavel. What are biomarkers? *Current opinion in HIV and AIDS*, 5(6):463–466, 2010.

[57] Kasper Thorsen, Karina D Sørensen, Anne Sofie Brems-Eskildsen, Charlotte Modin, Mette Gaustadnes, Anne-Mette K Hein, Mogens Kruhøffer, Søren Laurberg, Michael Borre, Kai Wang, et al. Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis. *Molecular & Cellular Proteomics*, 7(7):1214–1224, 2008.

[58] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics (Oxford, England)*, 25(9):1105–11, May 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp120.

[59] Natalie A Twine, Karolina Janitz, Marc R Wilkins, and Michal Janitz. Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by alzheimer's disease. *PLoS One*, 6(1):e16266, 2011.

[60] Mathias Uhlen, Per Oksvold, Linn Fagerberg, Emma Lundberg, Kalle Jonasson, Mattias Forsberg, Martin Zwahlen, Caroline Kampf, Kenneth Wester, Sophia Hober, Henrik Wernerus, Lisa Björling, and Fredrik Ponten. Towards a knowledge-based Human Protein Atlas., 2010.

[61] user:Rgocs / Wikipedia: The Free Encyclopedia. RNA-Seq alignment with intron-split short reads, 2009. URL `http://en.wikipedia.org/wiki/File:RNA-Seq-alignment.png`.

[62] Vladimir N. Vapnik. Statistical learning theory. *Wiley*, 1998.

[63] Kai Wang, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, Piotr Mieczkowski, Sara a Grimm, Charles M Perou, James N MacLeod, Derek Y Chiang, Jan F Prins, and Jinze Liu. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic acids research*, 38(18):e178, October 2010. ISSN 1362-4962. doi: 10.1093/nar/gkq622.

[64] Weichen Wang, Zhiyi Qin, Zhixing Feng, Xi Wang, and Xuegong Zhang. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*, 518(1): 164–70, April 2013. ISSN 1879-0038. doi: 10.1016/j.gene.2012.11.045.

[65] Yu Wang, Igor V Tetko, Mark A Hall, Eibe Frank, Axel Facius, Klaus F X Mayer, and Hans W Mewes. Gene selection from microarray data for cancer classification–a machine learning approach. *Computational Biology and Chemistry*, 29:37–46, 2005.

[66] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[67] Brian T Wilhelm and Josette-Renée Landry. RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing. *Methods (San Diego, Calif.)*, 48(3):249–57, July 2009. ISSN 1095-9130. doi: 10.1016/j.ymeth.2009.03.016.

[68] Xiaolin Xu, KaiChang Zhu, Feng Liu, Yue Wang, JianGuo Shen, Jizhong Jin, Zhong Wang, Lin Chen, Jiadong Li, and Min Xu. Identification of somatic mutations in human prostate cancer by RNA-Seq. *Gene*, 519(2):343–7, May 2013. ISSN 1879-0038. doi: 10.1016/j.gene.2013.01.046.

[69] Kevin Y Yip, Chao Cheng, and Mark Gerstein. Machine learning and genome annotation: a match meant to be? *Genome biology*, 14(5):1–10, 2013.

[70] Yanju Zhang, Eric-Wubbo Lameijer, Peter a C 't Hoen, Zemin Ning, P Eline Slagboom, and Kai Ye. PASSion: A pattern growth algorithm based pipeline for splice junction detection in paired-end RNA-Seq data. *Bioinformatics (Oxford, England)*, 28(4):479–486, January 2012. ISSN 1367-4811. doi: 10.1093/bioinformatics/btr712.

Vita Auctoris

Ahmad Tavakoli was born in 1987 in Mashhad, Iran. He graduate from the Ferdowsi University of Mashhad in 2009 with a Bachelor of Science degree in Computer Engineering. He joined the University of Windsor's School of Computer Science in September 2011 and earned his Master of Science degree in September 2013.